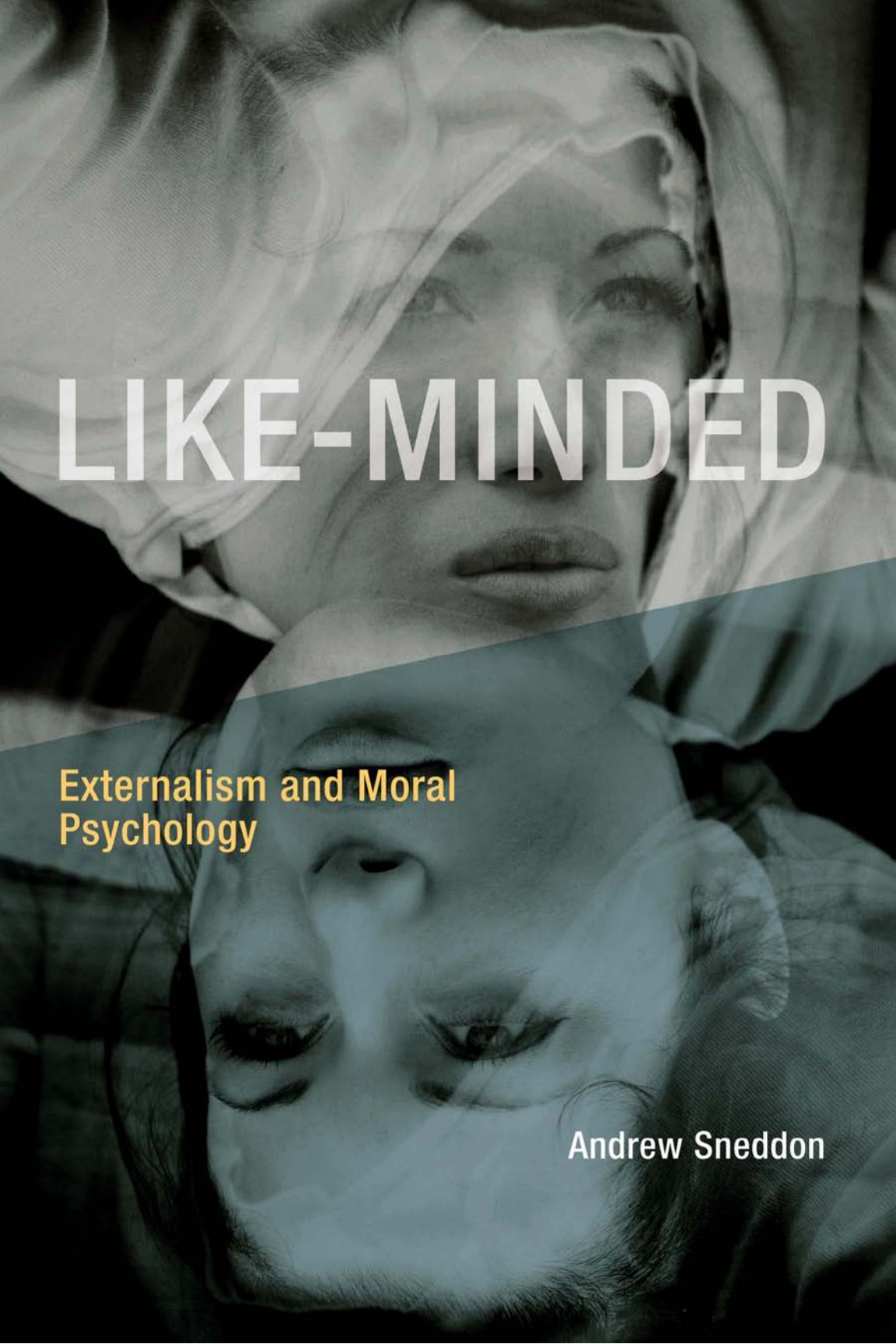

LIKE-MINDED
Externalism and Moral Psychology
Andrew Sneddon

Like-Minded

Like-Minded

Externalism and Moral Psychology

Andrew Sneddon

The MIT Press
Cambridge, Massachusetts
London, England

Set in Stone Sans and Stone Serif by Toppan Best-set Premedia Limited.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Sneddon, Andrew, 1971–
Like-minded : externalism and moral psychlogy / Andrew Sneddon.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-01611-7 (hardcover : alk. paper)
1. Ethics. 2. Psychology and philosophy. 3. Externalism (Philosophy of mind). I. Title.
BJ45.S63 2011
170'1—dc22

2011001922

10 9 8 7 6 5 4 3 2 1

for Debbie

Contents

Preface

My project in this book is to chart the boundaries of the psychology of moral agency. My method is to unite two discussions in philosophical psychology that have to date proceeded independently of each other. On one hand there is the booming interdisciplinary work done in philosophical moral psychology since the 1990s. In those years, when I was a student, this field was nascent. Now it is arguably the most active corner of both philosophical psychology and moral philosophy. Despite my criticisms of that work, I have been greatly impressed by the philosophers and psychologists who have jointly shed so much light on the psychological capacities that make us moral agents. On the other hand there is the philosophical debate about the role of agents' contexts in their minds. That debate, between "individualists" and "externalists," has its roots in work done in the 1970s on mental meaning, but since 1998 its focal point has broadened into what is often called the "Extended Mind Hypothesis." Individualists hold that an agent's context can provide input only to cognitive processes—i.e., contextual features are not parts of cognitive processes themselves. Externalists argue that features of an agent's context can be constitutive parts of cognitive systems, not just sources of input. I have become convinced that the individualism/externalism issue should be seen as an empirical one. If progress is to be made here, it will be made by designing both individualistic and externalistic hypotheses and testing their ability to explain psychological phenomena. Since philosophy and psychology tend toward individualism, relatively few externalistic hypotheses have been formulated and tested. I aim to fill this gap partially by presenting a generally externalistic position about human moral psychology. I call this the Wide Moral Systems Hypothesis.

It is composed of more particular hypotheses about moral judgment, moral reasoning, the attribution of moral responsibility, and the production of action.

I began this project in 2003 with a Standard Research Grant from the Social Sciences and Humanities Research Council of Canada, for which I have long been grateful. I would like to thank the Arts Faculty of the University of Ottawa for additional travel and research support. Jim Greenwood and Mark Young provided valuable research assistance during the early years of this project. Thanks go to the students in my 2007 moral psychology graduate seminar, who read early versions of some of this material. The manuscript was improved by the attentive comments made by the readers for the MIT Press, for which I am indebted. Thanks are also due to members of audiences at talks that I have presented on versions of this work at Carleton University, at the University of Ottawa, at York University, at the University of Montreal, at Washington State University, and at the College of the Holy Cross. Some of this work has appeared in journals. I am grateful for permission to use material from the following articles:

The depths and shallows of psychological externalism, *Philosophical Studies* 138 (2008), no. 3: 193–208

A social model of moral dumbfounding: Implications for studying moral reasoning and moral judgment, *Philosophical Psychology* 20 (2007), no. 6: 731–748

Two views of emotional perception, in *The Modularity of Emotions* (special supplement to *Canadian Journal of Philosophy*, 2006), edited by C. Tappolet and L. Faucher

Like-Minded is about the way context can be a part of our minds. Thus, I would be particularly remiss if I did not acknowledge the context in which the book was born. Having acknowledged my professional context, I should now acknowledge two more personal debts of gratitude. First, I spent a great deal of time thinking about this project and these topics while walking my dog through the streets of my Ottawa neighborhood. Certain blocks in Wellington Village and Westboro still evoke thoughts about moral judgment, moral responsibility, moral dumbfounding, and related topics for me. I find this very pleasant. I consider myself very

fortunate to live in such an enriching and enjoyable place. Second, my wife, Debbie, has known me since before I took my first course in philosophy. She has been with me as I have worked through these and other topics, sometimes fruitfully and sometimes pointlessly. During my work on the book, she has been a psychological subject, a philosophical colleague, and a beloved companion on many of those walks with our dog. This is for her.

1 Introduction: Externalism and Moral Psychology

1.1 Reasons, Passions, and a Third Option

What are the psychological foundations of morality? What psychological capacities enable us to evaluate actions? To act in accordance with moral norms? To attribute moral responsibility to ourselves and others? Although these have been perennial concerns for philosophers, there has been a flurry of work on them in recent years in a distinctly interdisciplinary vein. Philosophers and psychologists have combined resources to address these questions. The results include both new formulations of familiar positions and genuinely new answers. This book contributes to this interdisciplinary trend.

Historically the chief question for philosophers has been whether the psychological foundations of morality are emotional or rational. The classical protagonists in this debate are well known: David Hume (1740) argued that reason is the slave of the passions, so morality must be based on them, whereas Immanuel Kant (1785) argued that moral law is given by rational agents to themselves in virtue of their rationality. This debate continued through the development of analytic meta-ethics in the twentieth century, and it continues today. Simon Blackburn (1998) is a prominent intellectual descendant of Hume, while Michael Smith (1994, 2004) is arguably the most prominent present-day rationalist. Empirical data have been brought into this debate. For example, Shaun Nichols (2002, 2004a) has argued that empirical studies of psychopathy support a Humean view of morality rather than a Kantian one. Jeannette Kennett (2006) has recently defended moral rationalism from this charge on empirical grounds.

My primary aim is to make a third option plausible. The words 'reason' and 'passion' do not satisfactorily capture all of the important options for

explaining the psychological foundations of morality. A third possibility is, to put it roughly, that these foundations centrally include capacities that enable us to operate within cognitive systems that extend beyond individual agents into the wider world. I call this the *Wide Moral Systems Hypothesis*. This hypothesis fits within the array of positions known as *externalism* about the mind or, rather more catchily, the *Extended Mind Hypothesis*. According to the Extended Mind Hypothesis, at least some cognitive processes extend beyond the individual agent to include worldly resources. These resources are not merely input to cognitive processes that are located within an individual's brain. Rather, they partially constitute the cognitive processes in question. The conventional terminology is to call processes that are partly constituted by environmental resources "wide"—hence the name of the general hypothesis defended in this book. Processes that are located solely within the bounds of agents' bodies are "narrow."

Each of the next four chapters presents specific wide hypotheses about a distinct aspect of our moral psychology. The first topic is moral judgment. This is a traditionally central topic in examinations of moral agency. However, we must be careful with this term. For one thing, it is easy to assume that "judgment" must be the product of a process of judging, and that this in turn is something done consciously by an agent, perhaps analogously to what is done by a legal judge in a courtroom. I wish to avoid these assumptions. For another thing, confusion about this term has arisen. Following Jonathan Haidt (2001), Marc Hauser (2006), and Jesse Prinz (2006a), I use "moral judgment" here to refer to the psychological capacity or capacities by which we evaluate things actions, states of affairs, and persons in moral terms, however this is accomplished. Some people—e.g., Jorge Moll et al. (2005)—define moral judgment such that it is automatically the product of moral reasoning. Doing so eliminates any substantial inquiry into whether the foundation of moral judgment is moral reasoning. I follow many psychologists and philosophers in taking this as a substantial issue, one to be decided through conceptual and empirical inquiry rather than by definition. Accordingly, I do not define moral judgment as the product of moral reasoning.

Besides moral judgment, I will be examining moral reasoning, the production of action, and attributions of moral responsibility. Despite the

judgment-centric approach of most discussions of moral psychology, I am inclined to think that these phenomena are just as central to moral agency as moral judgment and worth just as much attention. Regardless of questions of priority, a good case can be made for thinking that the items on this list come close to exhausting the range of our central moral-psychological capacities. In chapter 6, I look around for topics to add to our view of moral psychology. Lots of work has been done on these topics in both philosophy and psychology, some of it now well known and considered classic. I am going to revisit this work with an eye on what I take to be its overlooked externalist aspects. Again and again I have been struck by the integration of agent and environment either described or hinted at by research on moral judgment, moral reasoning, moral motivation, and moral responsibility. There is a story both familiar and novel here, and I intend to tell it as best I can.

There is work to be done before I turn to moral psychology. In this chapter, I provide some conceptual tools for thinking about cognitive systems that extend beyond the physical boundaries of individual agents. This will illuminate the conceptual possibility of such systems. The bulk of the book will be concerned with establishing the empirical plausibility of the Wide Moral Systems Hypothesis.

1.2 The Extended Mind Hypothesis: Varieties of Individualism and Externalism

A closer look at the individualism/externalism debate is required in order to see just how the Wide Moral Systems Hypothesis is an alternative to the traditional options. Arguably, the reason/passion debate in moral psychology has been about psychological capacities that are attributed to individuals. This is the default approach in both empirical and philosophical psychology: questions are framed in terms of capacities attributed to individuals. However, over since about 1970 philosophical psychology has been marked by the sustained challenge to this approach that has come to be known as the Extended Mind Hypothesis. Defenders of this hypothesis are typically known as "externalists"; those who deny it are "individualists." In very general terms, the debate between individualists and externalists is about how to understand the role of an agent's context in

the agent's psychological functioning. Individualists restrict context to the psychological background. It is a source of input to our psychological processes and it receives output from them. Externalists do not deny that context performs these functions. However, externalists claim that contextual features can also be parts of psychological processes. To refine our sense of the issues, here are some central ways in which externalist theses have been refined and developed.

First Distinction: Content Externalism and Vehicle Externalism

Today's philosophical debate about externalism has its roots in philosophy of mind and language. The seminal thought experiments of Hilary Putnam (1975) and Tyler Burge (1979), with their emphasis on the meaning of utterances and the content of such folk psychological states as beliefs, exemplify this approach. In an assessment of the debates arising from these thought experiments, Mark Rowlands (2003) distinguishes *content* externalism from *vehicle* externalism.

Let's begin with content. The individualist about content holds that the meanings of utterances and mental states are logically independent of environment. The externalist denies this. Putnam and Burge's now-classic thought experiments work by probing our intuitions about what happens when we hold the intrinsic properties of individuals constant and vary their environments. Their claim is that what is revealed by such arguments is that mental and linguistic content turns out to vary with environmental variances despite the constancy of the agents' intrinsic properties.

Content externalism implies nothing about the nature of the items that bear content. Putnam, Burge, and subsequent individualists and externalists about content have not been primarily concerned about the nature of our cognitive architecture. It is perfectly consistent with thoroughgoing content externalism to hold that the bearers of psychological content are intrinsic features of agents. Debate about vehicle externalism calls this directly into question. Physicalist individualists about this issue hold that the vehicles of content are located within the physical bounds of individual agents. Vehicle externalists deny this. Since this issue is distinct from that of content externalism, new arguments are needed to assess the plausibility of these positions. And since the vehicle issue seems to be at least in part about the mechanics of psychological processes, empirical information is more relevant here than it is to content externalism.

Second Distinction: Taxonomic and Locational Externalism

The importance of empirical study to the assessment of vehicle externalism introduces a second way of classifying varieties of externalism. This is because the other way to approach this territory has been from the perspective of philosophical psychology and philosophy of science more generally. Robert Wilson (2003, 2004) takes this approach and distinguishes *locational* externalism from *taxonomic* externalism.

One way Wilson draws the locational/taxonomic distinction is by examining the metaphysics of the relation "realization." Following Sydney Shoemaker, Wilson makes consideration of systems the starting point of his case. For a higher-level property H and a system S in which it is realized, the *core* realization is "a state of the specific part of S that is most readily identifiable as playing a crucial causal role in producing or sustaining H" (Wilson 2001, 8). The *total* realization of H is "a state of S, containing any given core realization as a proper part, that is metaphysically sufficient for H" (ibid., 8). When a system is contained within an individual, an individualistic interpretation of the properties of that system is warranted. However, Wilson draws our attention to the possibility of systems that include individuals as a part and hence extend beyond the physical boundaries of individual agents. Using this possibility, Wilson identifies two sorts of externalist realization. "Wide" realization occurs when there is "a total realization of H whose non-core part is not located entirely within B, the individual who has H" (ibid., 11). "Radically wide" realization involves "a wide realization whose core part is not located entirely within B, the individual who has H" (ibid., 13).

When a property of an individual is widely realized, it must be individuated in reference to the system that extends beyond the boundaries of the individual. This yields a position about taxonomy: taxonomic externalism (Wilson 2003, 276; 2004, 174–178). In contrast, properties with radically wide realizations are not solely properties of the individual one is examining, but are instead located at least partly beyond its physical boundaries. The associated view of externalism is, accordingly, *locational* externalism (Wilson 2003, 276; 2004, 174–178).

These characterizations of externalism apply not only to psychology but to any phenomenon to which the metaphysics of realization of properties by systems applies. Wilson (2004, 114–115) draws examples from biology. The biological property of *being a predator* is one that is properly attributed

to individual organisms, but one that they have by virtue of their role in a predator-prey system. Accordingly, biologists should be taxonomically but not locationally externalist about predators. However, Wilson's central examples of locational externalism come from psychology. There are research programs in cognitive science that describe cognitive tasks as being accomplished between individuals, or via individual-environment interaction. Wilson discusses Edwin Hutchins's work on how navigational tasks are performed (1995) and Rodney Brooks's work designing robots (1991) as examples of such research programs. The contention is that the cognitive processes in question are located partially beyond the physical boundaries of the individuals participating in the systems, so we should be locationally externalist about them.

Let's return to the content/vehicle distinction. Taxonomic externalism is equivalent to content externalism only if principled scientific psychological taxonomy is done only in terms of the content of psychological states. Taxonomy by content is undeniably important. However, whether it is the only way scientific psychology can characterize the elements of its domain seems to be an open question. As a broad possibility, perhaps some items in psychological explanations ought to be individuated in functional terms, i.e., in terms of their relationships to input and output. If this is the case—and whether it is seems to be an empirical issue—then taxonomic externalism is distinct from content externalism.

In contrast, locational externalism is equivalent to vehicle externalism only if the only way in which the bearer of content can extend beyond the intrinsic boundaries of an individual is for it to be *realized* by a *system* of which the relevant individual is a part. If realization is not the only relation relevant to the nature of bearers of content, or if there are principled ways of addressing bearers of content independent of their possible or actual roles in systems, then locational externalism is distinct from vehicle externalism. Again, these appear to be empirical issues.

Putting these nuances aside, it is reasonable to see the content/vehicle and taxonomic/locational distinctions as deeply related. That said, Wilson's turn from the traditional concerns of philosophy of mind and language to empirically informed metaphysics is both a genuine step forward in the development of externalism and a minor obstacle to a clear view of the possible implications of externalism. It is a step forward in that it both acknowledges and makes clear the connection

between this issue and empirical work in various sciences, particularly psychology. It obscures matters because of its emphasis on the metaphysics of realization, which is at least one step removed from the practical concerns of practicing psychologists and opaque with regard to its relevance to these concerns.

The focus on the practice of psychology brings us to a distinction that, although implicit in content/vehicle distinctions and (especially) in taxonomic/locational distinctions, has gone undeveloped. For psychological externalism to be empirically assessed, psychological hypotheses must be framed in terms that are relevant to differences between individualistic and externalist interpretations of psychological phenomena. The most straightforward way for this to happen is for the hypotheses themselves to be framed in explicitly externalist terms. How are we to know what topics call for externalist hypotheses? I have no thoroughgoing answer to this question, but Wilson's work suggests a starting point. Externalist hypotheses are warranted for any psychological phenomenon that exhibits *systematic* individual-environment relations. I am inclined to think that the question of when an individual-environment system is present is one that must be answered *a posteriori*, and that particular sciences may justifiably have differing working notions of conditions that must be satisfied for the presence of a system. Nevertheless, Wilson's work provides us with a rough notion of what I shall call "systemicity" that can be used as a rule of thumb. (Note well: The following is not offered as an analysis of "system" in necessary and sufficient conditions.)

In refining the concepts central to Developmental Systems Theory, Wilson (2005, 153) characterizes developmental systems as follows: "Developmental systems must be causally and functionally integrated chains of developmental resources, and these, individually and collectively, must play a replicable causal role in ontogeny and inheritance." If we strip this of content peculiar to developmental systems, we have a general schema for systemicity:

_____ systems must be causally and functionally integrated chains of _____ resources, and these, individually and collectively, must play a replicable causal role in _____.

At present we are interested in particular kinds of psychological systems, so the resources in question must be *cognitive* ones, broadly understood as

informational input and output domains and mechanisms. To justify the description of something as a psychological system, these resources must play a replicable causal role in the production and the execution of particular psychological phenomena. To instantiate a *moral*-psychological system, there must be the appropriate sorts of cognitive resources connected in the appropriate ways to produce and sustain particular aspects of moral cognition.

The possibility of systems that are distributed between an individual and that individual's environment—i.e., of *wide* systems—is delivered by the possibility of the requisite causal and functional integration. The higher the degree of causal and functional integration there is between an individual and aspects of the individual's environment, the greater the reason there is to think that the individual and those aspects of the environment constitute a system. In the chapters that follow, our attention will be on exactly this sort of individual-environment integration with regard to our central moral-psychological capacities.

Using the notions of systems and systematic individual-environment interaction as our starting point, we can distinguish two forms that externalist hypotheses can take:

A psychological hypothesis is *shallowly* externalist when it begins with psychological items attributed to an individual regardless of environmental integration and construes them widely.

A psychological hypothesis is *deeply* externalist when it begins with systematic individual-environment interaction and attributes psychological items to the individual as needed to participate in the given wide system.[1]

The difference between shallow and deep externalism is one of initial presuppositions. Shallow externalist hypotheses are to be expected when they are framed as reinterpretations of individualistic hypotheses. In these cases, it is reasonable to interpret what one encounters as a relatively superficial modification of one's understanding of something with an individualistic basis. This is one way of understanding debates about content externalism: propositional attitudes are attributed to individuals, which individualists had construed as logically independent of context but which externalists reconstrue as logically dependent on certain contextual features. In both cases, exactly the same psychological items are attributed to

individuals. In contrast, deep externalist hypotheses are framed from an externalist starting point, rather than as a reinterpretation of previously individualistic ideas.

The traditional reason/passion debate is reasonably interpreted as consisting in the examination of either individualistic or shallowly externalist hypotheses. Reason and passion are understood as psychological items that can be attributed to individuals regardless of context. In contrast, the Wide Moral Systems Hypothesis (WMSH) is a deeply externalist position. My primary aim is to make plausible the idea that the psychological foundations of morality should be understood, at least partly, in terms of cognitive systems that extend into the environment beyond the physical bounds of individual agents. Psychological items will be attributed to individuals on the basis of such systematic agent-environment interaction, where it is found. In short, context is treated as integral to our central moral-psychological capacities in the WMSH.

Although the particular nature of the psychological items attributed to individuals in deeply externalist hypotheses must depend on the details of the case, two things can be said in general. First, these items are for the individual to participate in the wide system; they are not for replicating whatever psychological functions the wide system performs.[2] The idea is that some psychological job P is performed once by the wide system, not twice (once by the wide system *and* once by narrow systems that an individual happens also to have). Sometimes we will encounter cognitive redundancy, where P can be performed, and may even actually be performed, by both wide and narrow systems. However, there is no *a priori* reason to require individuals to narrowly perform P while participating in wide systems that also perform P. In fact, if this were taken to be a point about systems generally, there would be an *a priori* reason against it, as it would generate an infinite regress: for a system to perform P, there would have to be some other system to do so. For this second system to perform P, there would have to be a third system, *ad infinitum*. Second, the psychological items attributed to an individual to participate in a wide cognitive system need not themselves be wide in every sense. To be precise, they need not be locationally wide. They will be taxonomically wide, insofar as they have to be classified in terms of the wide system in which they play a role. But they themselves can, quite comfortably, be located within the

physical boundaries of individual agents. There is good reason to think that locationally narrow but taxonomically wide psychological capacities provide important underpinnings for our moral psychology.

1.3 What about Twin Earth?

Let us briefly return to content externalism and the famous arguments of Putnam and Burge. They employed a specific sort of argument that many will associate with discussion of externalism in general. This type of argument asks us to compare pairs of linguistic contexts. Crucially, we are to compare people in these contexts whose intrinsic, individualistic properties are identical. The contexts themselves differ in some specific way. For instance, Burge presents two people who are individualistically identical. Their contexts vary with regard to the meaning of the word 'arthritis'. In one context, the word applies only to certain disorders of joints; in the other, it also applies to disorders in body parts other than joints. When a person in the first context complains of arthritis in his thigh, he speaks incorrectly. When an identical person in the second context makes the same complaint, he is truly speaking of arthritis in his thigh (Burge 1979, 77–79). Putnam asks us to compare Earth and "Twin Earth," in the process giving rise to the tradition of referring to such arguments as Twin-Earth arguments. On Earth, water is H_2O, but on Twin Earth it has a different chemical constitution, which Putnam calls XYZ. When Oscar_1 on Earth uses the word 'water', he means H_2O even if he has no idea what elements constitute water. When Oscar_2—who is intrinsically identical to Oscar_1—uses the word 'water' on Twin Earth, he means XYZ even if he has no idea what the constituents of water are (Putnam 1975, 139–141). For both Burge and Putnam, the point is that mental and linguistic content are at least partly determined by the contexts in which agents find themselves. Agents' individualistically construed properties do not suffice to determine the content of their thoughts and their utterances.

Twin-Earth arguments are so closely tied to debates about externalism and individualism that some readers will expect to find some in this book. You will not find any. There are two reasons for this. The first is that, as we have just seen, Twin-Earth arguments concern, first and foremost, questions of content. Content is not the topic of this book, so the Twin-Earth considerations of Putnam and Burge find no natural application here.

Some may find this lame. Surely the construction of Twin-Earth arguments for topics other than content is not impossible. It takes patience, imagination, and hard work, not magic, meaning that their omission is suspicious. Let us admit for the purposes of argument that Twin-Earth arguments can be constructed for topics other than content. There is a second and more important reason for omitting this way of arguing in the present book. I suspect that Twin-Earth arguments are useful for developing and evaluating shallow externalism. However, I am impressed by the more radical possibilities offered by externalism.[3] This calls for deeply externalistic hypotheses, but Twin-Earth arguments are not useful tools for devising such hypotheses. Twin-Earth arguments work by holding some pre-specified feature or features of agents constant and varying the contexts in which the agents function. This invites a conservative approach to the description of agents. The reason is that there is a tendency toward individualism in both folk and scientific psychology. This bias results in the description of agents in terms that are usually used individualistically. In Twin-Earth arguments these descriptions are subsequently re-imagined widely. This is shallow externalism. If I am correct that the theoretical possibilities of externalism run deeper, then we will do well to avoid ways of arguing that invite shallowly externalistic hypotheses. Consequently there will be no more visits to Twin Earth in this book.

1.4 Objections to Psychological Externalism

Of course, there have been objections to the Extended Mind Hypothesis. The positions of Putnam and Burge have long been resisted. Indeed, Burge's seminal 1979 paper is largely constructed around potential objections to the very idea of externalism about mental content. In his 1995 book, Wilson scrutinizes two decades of work favoring individualism and finds it wanting. The more recent varieties of externalism have met equally persistent opposition. Frederick Adams and Kenneth Aizawa (2001, 2008) and Robert Rupert (2004, 2009) offer significant challenges. Important responses can be found in Clark 2008 and in Wilson and Clark 2008. I shall not delve into the details of the discussion about the objections made by Rupert and by Adams and Aizawa, since I think that the decisive responses have already been made. A more recent objection that has not

yet been adequately answered is that of Mark Sprevak (2009). Sprevak's case deserves some attention before we turn to empirical issues.

Sprevak's argument has two ideas at its core: the Parity Principle and the Martian Intuition. The Parity Principle, which Sprevak calls the "fair treatment" principle (2009, 505), comes from the famous formulation of externalism by Andy Clark and David Chalmers. The Parity Principle is designed to focus attention on our judgments of what systems, states, and processes are cognitive and to divert attention from putatively misleading side issues such as whether a system is located solely within the physical bounds of an organism:

> *The Parity Principle* If, as we confront some task, a part of the world functions as a process which, were it to go in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process. (Clark and Chalmers 1998, 8)

It should be clear that the specific issue is locational externalism. For present purposes, the Parity Principle provides a partial rule of thumb for deciding whether a process is cognitive (or mental—Clark and Chalmers address both). Whether it is the only idea relevant to such decisions is one of the things Sprevak examines with his argument.

The Martian Intuition is that it is conceivable for creatures with mental states to exist even if they are physically and biologically different from humans (Sprevak 2009, 507). This idea has long had a role in philosophy of mind: it has been a famous part of arguments leading from identity theories to functionalism (ibid., 509). Blood, skin, hearts, and the like are, *prima facie*, inessential to mentality. So perhaps are brains, spines, and nerves. According to functionalist deployments of the Martian Intuition, the important thing is what these substances do, not the substances themselves. This implies that creatures made of silicone or mud or tin cans (use your imagination) could have minds if these substances do the same things that neurons do for us. Such non-humans with minds are the "Martians" in question.

The Martian Intuition has three roles in Sprevak's argument. First, following Clark, Sprevak uses it to answer objections to locational externalism. Rupert (2004) and Adams and Aizawa (2008) object to externalism by appealing to fairly fine-grained features of putatively brain-bound or body-bound human cognition that extended processes do not share. Sprevak claims that arguments of this sort violate the Martian Intuition: it is

conceivable for creatures to have minds yet not to share the fine-grained features of human thought offered by the critics of externalism. Hence the objections turn on inessential features of mentality and are unduly chauvinistic.

The second role of the Martian Intuition in Sprevak's case is as the core of an argument for locational externalism from functionalism. Functionalism offers the functional organization of systems as the essential feature of mentality. The functional role of putatively mental phenomena must be specified to account for the nature of the phenomena. However, such specification can be done in myriad ways. Parameters must be provided to constrain such descriptions such that they provide all and only the relevant information. The Martian Intuition provides one of these parameters: functional roles must be specified in a sufficiently coarse-grained manner in order to allow for the possibility of minded creatures whose minds are realized in ways or substances different from the ways and substances that realize human minds. Sprevak argues that if the "grain parameter" is set at least coarse enough to allow for the possibility of Martian minds, then it will also allow for extended cognition in humans. The reason is that these cases of extended cognition will be as similar to brain-bound or body-bound human cognition as the Martian cognitive processes are. If we combine the Martian Intuition with the Parity Principle, then allowing for Martian minds but not extended human minds is unduly chauvinistic. Thus, if functionalism preserves the Martian Intuition, it also implies locational externalism.

Sprevak focuses on the version of externalism offered by Clark and Chalmers (1998). Besides functionalism, Clark and Chalmers specify three conditions that they think human-world processes must meet in order to count as cognitive. But, as with the objections of Rupert and those of Adams and Aizawa, Sprevak wields the Martian Intuition: these conditions are too fine-grained because we can imagine creatures with minds who do not share them. When combined with the Parity Principle, the implication of the cases generated by the Martian Intuition is that analogous cases which happen to extend into the world should count as cognitive. Hence, the constrained externalism of Clark and Chalmers is, again, unduly chauvinistic. Functionalism delivers *radical* externalism instead: *any* human interaction with a worldly resource suitable for use in a cognitive system constitutes an extended cognitive system. This is "radical" in that the constitution of

extended cognitive systems is unconstrained, and hence it is very easy for human-world systems to count as cognitive. Clark and Chalmers's constrained externalism is more modest because it sets limits on what sorts of systems can count as cognitive. Here is the third role of the Martian Intuition: radical externalism sets the bar of the mental so low that it allows phenomena that are, *prima facie*, non-mental to count as mental. For instance, by acquiring a book, the agent comes to believe everything contained in the book (Sprevak 2009, 517). The reason is that we can imagine Martians who encode beliefs using ink within the bounds of their bodies, are born with innate beliefs, and do not necessarily access these beliefs. The result is a Martian with the physical and functional equivalent of a book within its head which contains its non-accessed innate beliefs. If this system counts as mental, then, according to the Parity Principle, so should a system constituted by a person who has just acquired a book. Other examples: by stepping into a library I acquire millions of beliefs; by browsing the Internet I acquire billions of beliefs (Sprevak 2009, 518). Sprevak argues that, because it is implausible to count such processes as mental, radical externalism should be rejected. The Martian Intuition works against constraining externalism, so the problem can be traced back to functionalism itself. Externalism and functionalism probably contain insights into the mind and thereby provide useful material for developing their replacements (Sprevak 2009, 527), but if Sprevak is correct they are false.

The crux of Sprevak's critical argument is the construction of implausible cases of extended cognition using the combination of the Parity Principle and the Martian Intuition. However, the Parity Principle is subtler than the discussion so far suggests. As formulated by Clark and Chalmers and accepted by Sprevak, this principle functions by directing our attention to internal processes which we confidently count as cognitive and then extending this status, and the correlative confidence, to processes that extend beyond the bounds of agents into the wider world. But this is not the only place we find such confidence. We should see the familiar Parity Principle as the first part of a two-part principle. Here is the second part:

(PP2) If, as we confront some task, there is a process in the head that, were it to extend into the world beyond the agent, we would have no hesitation as accepting as not cognitive, then that process in the head is not a cognitive process.

We can be just as confident about what is not cognitive as we can about what is cognitive. PP2 codifies this confidence to guard against chauvinism in judging cases.

PP2 complicates Sprevak's argument. The reason is that the implications of our confident judgments about what processes count as respectively cognitive and not cognitive can be inconsistent. Consider Sprevak's confidence in the non-cognitive status of a system constituted by a person with a new book and his confidence in the cognitive status of the Martian processes constituted by ink-encoding of beliefs in the head, innate beliefs, and partial access to these beliefs. We have seen that Sprevak uses his confidence in the Martian case plus the Parity Principle to generate a judgment of "cognitive" for the person-plus-book system, which he finds to be unacceptably implausible. But we can run this argument the other way around. Let's begin with the person-plus-book system. Suppose that we confidently judge that this is a non-cognitive system. When this judgment is combined with PP2, the implication is that an analogous system that is located within the physical bounds of an agent should also be judged to be non-cognitive. The Martian system of ink-encoding of beliefs in the head, innate beliefs, and partial access to these beliefs is such an analogous system. Therefore we should see this system as non-cognitive. Strictly speaking, we should redescribe this case, insofar as 'belief' is a cognitive term that no longer applies here.

The ideas behind the Parity Principle turn out to be more nuanced than was expected, and the result in the present context is a clash of intuitions. Run one way, the Parity Principle generates a challenge to locational externalism and functionalism. Run another way, PP2 challenges our imagination about Martian cases. In general we could argue about which intuition is stronger, but this would not help Sprevak's argument, as he is committed to both the cognitive status of the Martian-ink case and the non-cognitive status of the book case. Moreover, such battles of intuitions are invariably unsatisfying. What is preferable is a principled way of adjudicating this clash.

Generally, and in a manner particularly germane in the present context, functionalism provides the tools for making progress with the (non-) cognitive status of these cases. The crucial question is whether these systems involve beliefs. Are there beliefs in the ink-Martian case? Does a person who buys a book thereby automatically come to believe the ideas

contained in the book by realizing a person-plus-book cognitive system? Functionalism directs us to things that are relevant to answering these questions. Here is Sprevak's characterization of functionalism: "Functionalism preserves the Martian intuition by claiming that what makes an organism have a mental state is the organism's functional organisation. This is typically understood in terms of the notion of a causal role, which in turn is understood as a pattern of typical causes and effects." (2009, 509) To assess whether our cases involve beliefs, we must have access to some specification of the typical causes and effects of beliefs. Too fine-grained a specification will render our account of belief chauvinistic. Too coarse-grained a specification will render our account too inclusive. Doing without any specification leaves us unable to make determinate judgments about cases. To see what might be in such a specification, consider two of the ways in which we can develop the book case:

(A) I buy a book. It is written in a language that I understand, but I have not read it yet. The topics are familiar to me, but I have not yet formed firm beliefs about them. If I were to read the book, I would be inclined to assert the ideas it contains. I would adjust my conduct in accordance with those ideas.

Here we have reason to think that the person-plus-book system really does involve something much like beliefs, albeit ones that are not yet deployed in on-line processing. The representations in question are accessible to my higher thought functions. If used on line, my earlier exposure to information about these topics would combine with these representations to deliver clear cases of beliefs. Potential access and responsiveness to rational processes informed by evidence are plausible hallmarks of belief. So are potential assent and suitability for guiding conduct. The pieces of a plausible characterization of the typical pattern of causes and effects of belief found here license the judgment that something much like beliefs, by functionalist standards, is found in this case. But now the case stands not as an implausible challenge to locational externalism, but as a plausible but surprising case of a cognitive system by functionalist standards.

(B) I buy a book. It is in a dead language for which no translating procedure exists, nor is there any reasonable hope of finding a "Rosetta Stone" key to its syntax and vocabulary. I have no familiarity with the topics. If

I were to read the book (which I cannot do), I would not be inclined to assent to its contents, nor would I adjust my conduct accordingly.

In this version of the case, virtually no elements of the pattern of the typical causes and effects of belief are found. Accordingly, we have no particular reason to see the person-plus-book system as a cognitive system involving beliefs. Again, this runs counter to Sprevak's argument: no implausible challenge to locational externalism is found here.

One might worry that the appeal to functionalism is question-begging in the present context, supposing that Sprevak claims that the cases generated by the Parity Principle and the Martian Intuition function as a challenge to functionalism. However, such a supposition would mistake the nature of Sprevak's argument. Sprevak argues from functionalism and the Parity Principle to the problematic cases. The present argument claims that Sprevak's argument omits important details about how the Parity Principle and functionalism work. That is, the argument claims that Sprevak's premises deserve more scrutiny. Once these details are taken into account, we see that Sprevak's argument does not go through. The problematic cases that are offered as a challenge to functionalism are not actually generated in the manner portrayed by Sprevak's argument.[4]

Let's put general objections to externalism behind us. From this point on, the important points will be more specific wide and narrow hypotheses concerning moral cognition. To prepare for these hypotheses, let's attend to wide systems themselves in more detail.

1.5 A Model for Thinking about Wide Cognitive Systems

A system takes inputs and delivers outputs, both of specific kinds. The relations between inputs and outputs are governed by if-then rules.[5] These rules need not be codified, of course; for some systems they are codified and for others they are not. For instance, I have an alphabetization-and-date system for my record collection: I take a musician's name as input, and the output is a particular place on my shelves for storing recordings by that musician. In the case of multiple recordings by the same musician, I take the date of recording as the input to a subsystem that also delivers a shelving order as output: earlier recordings are stored before later ones. I had never articulated these rules for this system until writing these words,

but in my case this particular system has existed for more than two decades.

Systems come in many kinds. Our present interest is in cognitive systems—that is, in the cognitive processes by which inputs are correlated with outputs. Moreover, since this book is an exercise in psychological theorizing about actual humans, the topic is causally realized cognitive systems that, by hypothesis, actually implement our thought about morality, rather than, e.g., merely formal systems. On paper my record-storing system is merely formal. When I put away batches of records, this system is causally efficacious in producing my actions. It is typical to think of cognitive systems as contained within the physical boundaries of individual people or other organisms. This assumption is challenged by the Extended Mind Hypothesis. Some wide systems will use cognitive resources, such as symbolically encoded symbols—e.g., printed letters and numbers. Other wide systems will use cognitive resources of other kinds. In the following chapters, I will argue that other people play a particularly important cognitive resource for moral psychology. To clarify the issues involved in thinking of psychological systems that exist between individuals, here is a simplified model.

Think of birds traveling in a flock (a tricky phenomenon to explain). Craig Reynolds (1987) has famously provided a computer simulation of the flocking of birds.[6] Reynolds calls his computer creatures "boids." Boids exhibit very realistic flocking behavior. This is achieved using three rules for steering:

Separation: Steer to avoid crowding local flockmates.

Alignment: Steer toward the average heading of local flockmates.

Cohesion: Steer to move toward the average position of local flockmates.

These rules require that boids monitor their immediate neighbors. The more complex phenomenon of flocking—that is, moving as a unit, dividing and recombining, and changing direction together—emerges from the behavior of individual boids as they track their local circumstances. They do not have a plan to form a group, or to follow a specific leader. This is very suggestive about how actual birds might accomplish their complex flocking behavior.

Now imagine a group of birds that act in accordance with the steering system codified above for boids. Actual birds do things other than fly

together—for example, they also seek food and avoid predators.[7] Let's add the cognitive capacities to find food and predators to our imaginary birds; for present purposes these need not be specified in any detail. Imagine the flock traveling through the air. The east-most bird sees food on the ground. This food is hidden from the west-most bird. The east-most bird heads toward the food. Nearby birds adjust their behavior in response to both the east-most bird and the food and subsequently follow. The west-most bird, in accordance with the three steering rules, adjusts its motion to keep up with the flock. As a result, the west-most bird ends up at the source of food.

Let's suppose, as seems quite plausible, that the behavior of the east-most bird can be explained in terms of psychological capacities located completely within that bird's physical boundaries. It takes the information about food as input, and the output is flying toward the food. The relevant systems are, by hypothesis, locationally narrow. How should we understand the behavior of the west-most bird? One possibility is that the behavior should be understood solely as a local response to the movements of its neighbors. Another possibility is that we should construe the bird's behavior as a response to the food and also as a response to its neighbors. One might balk at this interpretation on the grounds that the bird did not actually encounter the food, and so its behavior could not be a response to the food. However, externalist ideas give us a way to make sense of this: Perhaps the bird is part of a wide cognitive system. The input to the system is the information about the food. This information is taken in by, primarily, the east-most bird, which produces the output of turning toward the food. This information is taken as input by the intermediary birds and is subsequently processed via their responding movements, until it can produce the flying behavior of the west-most bird. Note that it is not required that the west-most bird *realize* that there is food to be had, or *know* about the food, or anything of the sort. To require that would be to suppose that the information about the food would have to be taken explicitly as input by the west-most bird in order for it to play a role in producing the bird's behavior. But such is not the case: the wide system, not the west-most bird, processes the information about the food. The west-most bird need only be able to play a role in this system of a sort suited for the processed input to produce the relevant behavior.

Thus we have two interpretations of the behavior of the west-most bird. How should we decide between them? One pertinent question, if not the

crucial question, is whether we are warranted in seeing the relations between the birds as systematic. To answer this, recall the schema for systemicity extracted from Wilson's work:

_____ systems must be causally and functionally integrated chains of _____ resources, and these, individually and collectively, must play a replicable causal role in _____

The issue of a "replicable causal role" can be put aside if we assume that this overall phenomenon is typical flocking behavior. The resources in question are, in the specific case, the information about the food, the steering capacities of the birds, the birds' movements that the steering capacities track, and the birds' food-detection capacities. What should we say about the causal and functional integration of these resources? Let's begin with causal integration. By hypothesis, the birds have specific steering rules for tracking their neighbors and responding to their whereabouts. That is, the behaviors typical of flocking are not by-products of more general procedures for moving or for tracking features of the environment. Thus, it seems to me that we are warranted in seeing the birds as exhibiting the requisite degree of causal integration. (N.B.: 'Degree' is the correct word here, as there is no specific line that, once crossed, divides systemicity from non-systemicity.)

These remarks about causal integration make reference to what the cognitive capacities in question are for—that is, they raise the issue of functional integration. In the case of boids, we can say that they exhibit functional integration because they were deliberately designed by humans to track each other in specific ways. Our imaginary birds are importantly different, insofar as nobody designed them. In this case the question of functional integration has to be addressed from a thoroughly naturalistic perspective. The natural way to address this, if not the only way, is to ask about the evolution of the birds' cognitive capacities. Without getting into the complex debate about the nature of natural functions,[8] here is a suggestion: if the finding of food via the following of nearby birds has contributed to the cross-generational persistence of the steering capacities by increasing reproductive fitness, then we have reason to think that flying in groups is not the only function of these capacities, and that finding food is another of their functions. By extension, birds' movements transmit information not only about themselves but also about the location of

food in the wider world.[9] Evidence about the evolutionary descent of such capacities might be difficult to gather, but the conjecture that finding food has contributed to the persistence of birds' navigational capacities strikes me as initially quite plausible, so I think we have *prima facie* reason to see the birds' movements, steering capacities, and environmental opportunities as functionally integrated to the requisite degree. That is, in this case we have reason to think that we find wide systemicity.[10]

Two *very* general things can be said about cognitive prerequisites for participating in wide systems, at least with regard to humans. First, the birds in this hypothetical case have the capacities to track the movements of their neighbors, but arguably other capacities are implicit in this example. The birds we have been considering are a sociable bunch: they are content to be around each other, and no conspecific hostility features in the case. This general state of affairs is important for participation in at least some wide systems that involve the use of some organisms as cognitive resources by other organisms. If the west-most bird were unwilling to pay attention to its neighbors, it would not be able to participate in the information processing that they make available. Second, although I have just described the birds as sociable, they are not nearly as social as humans.[11] The birds in this example participate in a wide cognitive system by tracking movements. In contrast, and in the spirit of much research into human sociality in general, I conjecture that many important wide systems in which humans participate require that we track each other's thoughts. If this is correct, then so-called mind-reading capacities are going to be required for individual humans to get access to wide cognitive resources.[12]

Let's return to the birds. If the wide interpretation is correct, then the west-most bird processes both information about the location of its neighbors and information about the location of food; it is aware, at most, of only the former. Thinking of the birds as taking part in a wide cognitive system allows us to distinguish three types of potential input. First, there is input to the individual's psychological capacities alone. Let's call this *unmediated* input. We can presume, for now, that this is an apt way to characterize the east-most bird's encounter with the food. Second, there is input that an individual does not directly encounter at all, but that is processed by the wide system. Let's call this *mediated* input. The west-most bird's processing of the information about the location of the food is mediated. Finally, there is information that is processed both directly by an

individual and by the wide system. Let's call this *dual* input. When an intermediary bird both sees the food and responds to the movements of the east-most bird, which is moving toward the food, it is dealing with dual input.

Dual input is tricky. It should give us pause with regard to how we think of unmediated input, at least for social creatures such as ourselves. Real birds not only follow each other and share food; they also compete for food and other opportunities. Humans are no different. But human social life is massively complex precisely because of the opportunities for manipulating each other for physical and social gain.[13] Thus, when two or more humans share input, we should be careful to include cognitive capacities for assessing and dealing with competition in our account of the psychological processes at work. This idea has at least two general implications. First, it makes the general openness to wide systems and resources an even more important prerequisite. Suppose that another person or some other organism is competing with me for food, status, and other opportunities, and that that individual has ideas about what it deserves, what its status is and should be, and what it can do to protect itself and to get ahead. It will be important for our generally agreeable co-existence that I appear to the other individual to have largely the same ideas. If I have different ideas—for example, that I, rather than he, she, or it, deserve X, and that my status is more important—then I pose a significant threat.

So far this point has been made in terms of two individuals and a limited number of topics of thought. But human life is far more complex than that. We interact with vast numbers of people, and about a relatively open-ended group of topics. We stand to each other in complex relations of power, status, threat, entitlement, and opportunity. Thus, it is not only important that I appear to agree with individual A about topic P; it is important that I generally fit in with most people, about an open-ended number of issues. This imposes on individuals a general, complex pressure to conform. This should set us up in particularly good position to realize wide cognitive systems with other individuals.

Consider how conformity might be psychologically implemented. Suppose that, regardless of how one actually thinks, there is reason to appear to agree with the views of others. One way to do that is to have mechanisms that suppress one's own contrary judgments and produce conforming behavior. But another way is to have mechanisms that conform

one's judgments to those of others. I see no reason to think that we do not have both sorts of mechanisms. If this is right, then thinking about competition and dual input has implications for how we think about the processing of unmediated input. If we have judgment-conforming processes, then the absence of other people in a particular situation is not significant: the effects of conformity extend all the way in, so to speak. Research into psychological heuristics provides some support for just such a phenomenon—see, for example, Gigerenzer 2008, 24; Richerson and Boyd 2005. Thus, even when dealing with input that is isolated from other individuals, it should be predicted that we will act as social animals. The availability (or the unavailability) of wide cognitive systems will be relevant even to the processing of unmediated input.

In light of the above, the term I use to describe people is *Like-Minded*. We are like-minded in two respects. First, as social animals we are under significant pressure to conform our views of the world to those of our conspecifics. Second, insofar as we participate in wide cognitive systems, partly in virtue of the psychology of conformity, there is an important sense in which we literally share psychological processes with other people. We think the same partly because it is prudent and partly because we use the same token systems to think. Sometimes we enter these systems as autonomous equals, sharing information through dialog and reasoning together to form judgments, solve problems, and generally figure things out. At other times, these systems are constituted not by explicit intersubjective reasoning but in other, less obvious ways. This is one of the lessons of the hypothetical birds: although they do not reason together explicitly, they nonetheless think together via subtler wide cognitive systems. The fact that these systems are relatively inconspicuous helps to explain why philosophers and psychologists tend to overlook them. Nevertheless, I am inclined to think that these less obvious ways of thinking together are the more important ones.

2 The Disunity of Moral Judgment

The topic of this chapter is moral judgment. By that I mean the psychological capacity or capacities by which we evaluate actions, states of affairs, and persons in moral terms. This is the central topic for most present-day philosophical moral psychology. I say "capacity or capacities" because of the question as to whether moral judgment is realized by one system or by more than one. Fiery Cushman and Liane Young (2009, 10) claim that in present-day moral psychology there is widespread agreement that moral judgment is realized by multiple systems. I agree with this view. In this chapter, I make a case for *plural*, *hybrid*, *embedded* moral judgment. Here are three brief stories to convey a sense of the contours of the view of moral judgment that will be defended in this chapter. They are accompanied by three questions. The answers to these questions show roughly how the present view differs from accounts of moral judgment offered by other philosophers and psychologists.

First Vignette: David and Immanuel are arguing about moral judgment. David thinks that emotions are the sole source of moral judgments and that reason does such other things as tracking relations among ideas, objects, and events. Immanuel thinks the opposite: true moral judgments stem from reason. Our emotions are pushed around by circumstances like leaves in the wind, so they are not suited to produce moral judgments.

Question 1: Can both David and Immanuel be correct?

This story raises the issue of pluralism about moral judgment. Despite the contention of Cushman and Young that there is widespread agreement about the multiplicity of systems for producing moral judgments, there is no shortage of philosophers and psychologists who think that moral judgment is psychologically unified. Those theorists must give a negative answer to the first question. In contrast, I shall argue that not only reason

and emotion but also other psychological phenomena can produce moral judgments. If we drop the insistence on "sole" or "true" sources of moral judgment, both David and Immanuel can be correct.

Second Vignette: Mary is visiting her friend June. They go for a walk in June's neighborhood and encounter a Caucasian man holding hands with an Asian woman. Mary has never seen a mixed-race couple before. June is aghast at the couple. She proclaims such love and behavior to be morally wrong. Mary has not thought about it before, but she is impressed by June's repugnance and agrees.

Question 2: How do June's emotions affect Mary's moral judgment?

This story raises the issue of externalism about moral judgment. Most theorists are individualists, and hence will tend to think that June's emotions can only provide input to Mary's moral-judgment-producing systems. On an individualistic view, psychological processes must be located within organisms' bodies. Since Mary and June are distinct individuals, their psychological processes must be distinct from each other. Externalists reject the idea that psychological processes must be located within the boundaries of particular bodies. On an externalist view, although June's emotions can provide input to Mary's moral-judgment systems, they can also play a more intimate role. I shall argue that one person's emotion can be a part of another person's capacity for making moral judgments. This is the sense in which moral judgment is "hybrid": the mechanisms of moral judgment are hypothesized to be constituted by systems composed of individual people and by features of the world beyond the physical bounds of these people. Some of the mechanisms of moral judgment are individual-world hybrids.

Third Vignette: While walking down the street, John sees an elderly woman's grocery bag break, dropping her food on the sidewalk. John instantly helps her, seeing that this is the thing to do.

Question 3: How intimately is John's moral judgment related to his action?

As we will see, the answer implicitly given by typical accounts of moral judgment is that, although moral judgment is important to the production of action, these are psychologically distinct phenomena. I reject this idea. Instead I shall argue for a more intimate relationship. The mechanisms of moral judgment are embedded in the sense that at least some moral judgments are psychologically indistinct from other psychological capacities,

such as moral reasoning and the production of action. John's judgment can be a psychological part of the processes that give rise to his action.

I shall begin with pluralism. Although I think Cushman and Young are correct about the multiple-systems view of moral judgment, I think they overstate the degree of agreement about this point. As examples of prominent theories of moral judgment I shall examine the works of Shaun Nichols, Marc Hauser, Jesse Prinz, and Jonathan Haidt. These theorists all present moral judgment as, to greater or lesser degree, fundamentally unified. That is, although they recognize various ways in which we evaluate actions, persons, etc., they all represent moral judgment as stemming from a single psychological core. The work of these theorists will be my principal foil in this chapter. In contrast, I think we should take the variety of ways in which we evaluate things more radically. I shall argue that the surface disunity of moral judgment goes all the way down to the root.

2.1 Unity Theories of Moral Judgment I: Shaun Nichols on "Sentimental Rules"

On Nichols's (2004a, chapter 1) "sentimental rules" theory, "core moral judgment" is realized by two mechanisms: a normative theory prohibiting harm to others and an affective mechanism activated by suffering in others. The combination of these mechanisms yields what Nichols calls "sentimental rules": the normative theory has rules prohibiting actions that activate, or are likely to activate, the affective mechanism. Moreover, the combination of these mechanisms is offered to account for our early abilities to distinguish moral and conventional transgressions: the affective mechanism is activated by a particular kind of badness (harm, rather than the breaking of rules of social interaction), and the normative theory encodes this distinctive reaction in rules whose status differs from that of rules about conventional issues.[1] These mechanisms constitute "core" moral judgment: Nichols (2004a, 90–96) argues that they appear early in our development, thereby realizing our earliest ability to perform moral judgments, and that they continue to function as the foundation of our abilities to make moral judgments when we are mature. An important part of Nichols's case for the role of sentimental-rules mechanisms in mature moral judgment is his examination of psychopathy (2004a, 17–20 and chapter 3; see also Nichols 2002). Psychopaths do not distinguish between

moral and conventional transgressions in the same way as normal people (Blair 1995; Blair et al. 2005, 58–59; Nichols 2002; 2004a, chapter 3). Nichols argues that the sentimental-rules mechanisms account for our ability to draw the moral/conventional distinction, and hence that specific dysfunctions in this mechanism can account for psychopathy.

After presenting the basic structure of his account of "core" moral judgment, Nichols points out that this structure is compatible with a range of more specific accounts of the architecture of moral judgment (2004a, 25–29). He suggests a spectrum of accounts: one pole gives the sentimental-rules mechanisms a role in the development of our moral psychology only; the other focuses on the on-line processing of these mechanisms in mature moral judgment only. In between are myriad hybrid possibilities. By casting his account as the basic structure of "a rich and complex phenomenon" (2004a, 26), Nichols acknowledges the variety of (at least possible) ways in which we make moral judgments. His overall position is that this variety is deeply unified by the sentimental-rules mechanisms.

2.2 Unity Theories of Moral Judgment II: Marc Hauser's Linguistic Analogy

On the basis of his analogy between moral judgment and our linguistic capacities, Marc Hauser hypothesizes that we have a "moral instinct." (See, e.g., Hauser 2006, 32–42.) Since remarks by John Rawls, influenced by Noam Chomsky's work in linguistics, foreshadowed Hauser's work, Hauser calls creatures with this sort of instinctual capacity for moral judgment "Rawlsian." He contrasts Rawlsian creatures with "Kantian" ones (whose moral judgments are produced by reason) and "Humean" ones (whose emotions produce moral judgments). Hauser does not think we are Humean creatures. That is, in contrast with Nichols, he does not think that emotion has a fundamental role in moral judging. He does think that reason can produce moral judgments, but he thinks this is a derivative capacity that comes relatively late in our development. Moral instincts are our earliest and most fundamental source of moral judgment. Like language, these instincts are constituted by principles to which we do not have introspective, first-person access.[2] These principles constrain possible moralities for humans who have them; cultural contexts set the specific settings of our moral judgments in various ways, all within these instinctual constraints.

Our Rawlsian instincts produce automatic, rapid judgments of actions on the basis of information about their causes and consequences; these judgments are about whether an action is permissible, obligatory, or forbidden (Hauser 2006, 46–55). Our Rawlsian instincts developed under distant evolutionary conditions, whereas our reasoning capacities are subject to more current influence. Thus, "[t]he Rawlsian creature will . . . fire off its intuitions about moral rights and wrongs, the Kantian will fire back principled arguments about these intuitions, and sometimes caught in the middle will be the Humean, generating angst, attempting to tilt the weight of the evidence toward one of the moral poles" (ibid., 418).

Hauser acknowledges variety in moral judgment in two ways. First, reason sometimes plays a role in generating moral judgments. Second, there are different moral codes, and hence there is the possibility of different moral judgments about the same situation (owing to contextual influences on the process that finely calibrates our capacities for moral judgment). However, in Hauser's account both of these rest on the platform of the inaccessible principles of the rapid, automatic, instinctual moral-judgment faculty. This faculty provides the parameters within which cultural influence operates, just as in the case of language. Reason operates on the basis of the more primitive action analysis that the instinct performs (Hauser 2006, 156–157). The moral instinct is the unifying feature of Hauser's account of moral judgment.

2.3 Unity Theories of Moral Judgment III: Jonathan Haidt's Social Intuitionism

Like Hauser, Jonathan Haidt (2001, 818–820) emphasizes the rapidity and automaticity of moral judgment. This leads him to think that reason does not play a fundamental role. Haidt casts reason as a process of conscious, intentional transformation of information. This is too slow and too controlled to account for the automatic features of moral judgment. Haidt calls this sort of cognition "intuition" (ibid., 818; see this page for a table of key differences between reason and intuition). Like Nichols but unlike Hauser, Haidt argues that intuitive processes include emotions (ibid., 814). The social aspect of Haidt's social intuitionism pertains primarily to the role of reason: Haidt thinks moral reasoning processes are primarily interpersonal, happening as between-agent exchanges of rationales for the

automatic judgments produced by intuition (ibid., 814; see 815 for a diagram of the structure of Haidt's account).[3]

Haidt wavers on the role of reason in moral judgment. Early in his flagship discussion, Haidt claims that "moral reasoning is rarely the direct cause of moral judgment (ibid., 815), which suggests that it can play this role. In his presentation of the details of his position, however, he relegates reason completely to derivative processing. If we follow Haidt's first suggestion, then his position is weakly unified: moral judgment typically happens one way, but can, occasionally, be produced in a distinct way. If we follow his second suggestion, then his position is thoroughly unified: there is only one sort of source for moral judgment.

So far I have emphasized the source of moral judgment in the social-intuition model. However, Haidt emphasizes the fluidity of moral judgment to a greater degree than Nichols and Hauser: "In the social intuitionist view, moral judgment is not just a single act that occurs in a single person's mind but is an ongoing process, often spread out over time and over multiple people." (2001, 828) I shall return to the externalist flavor of this view later in the chapter. For now I wish to emphasize the way that this builds heterogeneity into the social intuitionist account of moral judgment. If this view of the social intuitionist model is emphasized, Haidt's view is the least unified of the four Unity accounts of moral judgment. However, this way of interpreting Haidt conflicts with his early emphasis on the role of intuition in moral judgment.

2.4 Unity Theories of Moral Judgment IV: Jesse Prinz's Moral-Sensibility Theory

More than the previous three theorists, Jesse Prinz places emotions at the heart of moral judgment. For Prinz, a "sensibility" is a disposition to feel emotions of certain sorts. A moral sensibility is a disposition "to feel emotions in the approbation and disapprobation range" (2007, 92). Examples of emotions in the disapprobation range are contempt, anger, guilt, and shame (ibid., 79); the approbation range includes admiration, gratitude, and dignity (ibid., 81).

According to Prinz, moral judgment is constituted by the tying of one of these emotions to, e.g., an action. The action in question is categorized in a particular way. Prinz uses the examples of seeing an act of pick-

pocketing and classifying it as "stealing" (2007, 96–97). To make a moral judgment about this, one must have a moral sentiment about it. In this case, one must be disposed to have emotions about stealing that are in either the disapprobation range or the approbation range.[4] Once your sentiment is activated, the context plays an intimate role in determining what particular emotion is activated. For example, if you are disposed to disapprove of stealing, and you are not involved in the act in question, you might feel contempt at the pickpocketer's conduct. If you are the victim, it might instead be anger that is elicited. And if you are the thief, then perhaps you feel shame rather than an other-directed emotion. For Prinz (2007, 96), the association of the elicited emotion with the particular action is the moral judgment: "The compound [anger at pickpocketing] constitutes the judgment that pickpocketing is wrong, because the emotion that it contains was generated from a moral sentiment."

Prinz's account provides room for people to make different moral judgments from each other because of differences in their emotional dispositions. These dispositions are subject to influence from many sources, such as one's upbringing and broader social setting. However, in terms of psychology, Prinz's position is very much a Unity theory: the process I have so far outlined is the only way in which moral judgments are produced. Reasoning, for instance, influences moral judgment at the categorization stage (Prinz 2007, 122–125). It is not a distinct psychological source of moral judgment. Nor are the public processes of interpersonal exchange emphasized by Haidt to be seen as distinct roots of moral judgment. Such public expressions are "verbalizations" of moral judgments produced by the categorization-sentiment-emotion process.

2.5 Reflections on "Core" Moral Judgment

Let us now consider Nichols's term "core" moral judgment, since Nichols is the philosopher most explicit about defending what I am calling a Unity position regarding moral judgment. We can distinguish two senses of "core" relevant to this territory:

Developmentally "core" A feature that is developmentally "core" to ability X is one that is a necessary precursor for subsequent normal development of X.

Performatively "core" A feature that is performatively "core" to ability X is one that accounts for normal performance of X.

Nichols certainly means for the sentimental-rules mechanisms to be "core" in the developmental sense: they appear early and are necessary for subsequent development of other aspects of ordinary moral judgment. He wavers on whether they are performatively "core," although he tends to think that they are. When articulating the sentimental-rules position, he notes that there are various ways in which these mechanisms could figure in our actual judgments, and that assessing these possibilities requires more data (2004a, 25–29). Later, however, Nichols presents considerations suggesting that they are implicated in the performance of mature moral judgment. One of the models for this view is work on folk physics and folk psychology "according to which the 'core knowledge' of folk physics and folk psychology emerges early. Although the core knowledge might be 'enriched' through development, the early core knowledge is thought to persist unrevised into adulthood, and to continue to guide adult judgment. . . ." (Nichols 2004a, 93) Nichols criticizes varieties of sentimentalism that account for moral judgment in terms of complex psychological capacities that emerge later than our abilities to draw the moral/conventional distinction (ibid., 90–96). Overall, it is reasonable to think that Nichols contends that there is a psychological core to moral judgment that runs from our very early years into maturity, and that positions which obscure this cannot be correct.

It should be clear that we cannot infer that the early appearance of an ability in development implies that this ability is central to mature performance of related abilities. Developmental centrality need not imply performative centrality. Still, Nichols's emphasis on children's abilities and on the continuities between early developmental stages and mature moral judgment is worth taking to heart. That said, there is a psychological possibility that Nichols seems to have missed. On one hand, Nichols offers his position, which posits a continuous psychology of "core" moral judgment from early development into maturity. On the other hand, he criticizes alternative accounts that posit discontinuities, such that the mature performative core of moral judgment is not the same as its developmental core. For instance, he argues against a defense of neosentimentalism that has this structure. This view casts the capacity to judge the appropriateness

of guilt as the performative core of moral judgment (2004a, 93–94). What is overlooked is the possibility of *plural* mature moral-judgment abilities, without a core. The group of capacities that realize mature moral judgment can include those that appear early in development, but it need not rely on these features as performatively central. Instead, they can be one moral-judgment capacity among others. Such a position offers developmental continuity but no "core" moral-judgment capacity.

Is there any evidence to support such a view of mature moral judgment? Some details will have to wait for the discussions of autism and psychopathy in chapter 6. Disorders characterized by impairments of some ways of performing moral judgments but in which other ways of making such judgments remain intact provide good evidence for a heterogeneous psychological foundation of moral judgment. In the meantime, the heterogeneity of the theories offered by Hauser and Haidt is worth briefly noting, along with the evidence they use to come to their theories. Both Hauser and Haidt go to some pains to try to accommodate variety in moral codes with theories designed around a psychological core. Nichols also notes this variety. For instance, he remarks that, whereas our early abilities to draw the moral/conventional distinction are relatively stable across cultures, "more sophisticated forms of moral judgment are not cross-culturally stable" (2004a, 93). On the face of it, this combination supports a heterogeneous account of mature moral judgment at least as much as it supports a unified account.

So far, the topic has been *psychological* theorizing and *psychological* reasons for devising either a unified or a heterogeneous theory or moral judgment. However, more distinctly philosophical considerations are relevant to this topic. Consider *performative* centrality: a psychological capacity is the performative core of mature moral judgment if it is central to the normal making of moral judgments. The important thing to note here is that what counts as "normal" (or "core" or "central") moral judgment is partly a moral issue. In view of this, the question of what counts as "core" moral judgment cannot be settled by psychological considerations alone. However, Nichols and the others advance psychological considerations alone when devising their unified accounts of moral judgment. For an example, see Nichols's discussion of neosentimentalism (2004a, 93–94). Nichols focuses on whether "core" moral judgment as found in studies of the moral/conventional distinction is genuine moral judgment. Nichols

treats this as a thoroughly psychological issue, to be determined on psychological grounds alone. My contention is that this is, in an important respect, a moral issue, which implies that psychological considerations alone are not adequate to settle it. For another example, see chapter 7 of Hauser's *Moral Minds* (2006). Hauser begins the chapter by noting how versions of the Golden Rule have appeared throughout human history in diverse cultures. The spirit of this rule is to think of and to treat others by standards that one would like to have applied to oneself. The rest of Hauser's chapter examines cross-species studies of the psychology of cooperation and, more generally, living together. Hauser treats the apparent omnipresence of something like the Golden Rule as indicating a unifying feature of our moral psychology, deserving special focus—hence the dedication of a chapter to it. But he offers no normative argument for thinking that anything like the Golden Rule deserves centrality in our accounts of moral judgment. His argument is conducted on purely descriptive grounds.

To see what is problematic about relying solely on psychological considerations when theorizing about moral judgment, consider two contrasts. First, consider an intrapersonal contrast through time. Imagine a person who, as a young adult, values moral judgments that are made "from the gut." That is, this person trusts the moral judgments that come to mind first when presented by some scenario or information. Moral judgments made after the fact, after exposure to other sorts of information and reflection on the things that have been experienced, are treated as second-best. Now consider the same person decades later. Having given up the impetuous view of youth, the person now values moral judgments that are made on the basis of cool reflection. The first thoughts that come to mind when presented by a scenario or with information about some moral issue figure in the process of coming to a good moral judgment, but the person now sees these thoughts as second-best, to be discounted if careful thought about the relevant issues delivers verdicts that conflict with them. I take it that both ways of producing moral judgments are psychological possibilities for mature adults. The question for unity theorists is "Which is performatively (more) central?" It should be clear that this question cannot be answered on psychological grounds alone. From the youthful perspective, "gut" reactions are performatively "core" because this is the sort of thought that is valued as characteristic of the best kind of moral judgment. From the later-in-life perspective, "gut" reactions are of secondary impor-

tance. Instead, cool reflection is valued as the heart of mature, competent moral judgment. If a theorist places the automatic responses at the core of a theory of moral judgment on psychological grounds, this is subject to criticism on the grounds that the theorist has misunderstood the nature of *morality*, or what it is to be a *good* moral agent, or what we *ought* to value and strive for in moral judgment. The same criticisms can be aimed at theories that place reflective capacities at the core of moral judgment. These criticisms may be answerable, but they must be answered in moral terms, not psychological ones.

This contrast can be writ larger. Consider two groups, either within a common cultural setting or separated by cultural boundaries. Suppose that rapid, non-reflective ways of making moral judgments are valued in one group, and that cool, slow, reasoned ways of making moral judgments are valued in the other group. On psychological grounds alone, there is no way to decide that one of these ways of making moral judgments is performatively "core" and the other is not. Which sort of ability to put at the core of an account of moral judgment is partly an evaluative issue.

These considerations are wide-ranging. They suggest that *any* psychological theory that casts a particular way of making moral judgments as central to our overall capacity to make moral judgments can be met with a moral argument that this is mistaken. Insofar as the choice is not defended on moral grounds, and insofar as the criticism is reasonable, such a unity theory will be unsupported. The theories of Nichols, Hauser, Prinz, and Haidt all fit this mold. However, it is important to emphasize that these considerations do not apply to all attempts to articulate a core to all our capacities to make moral judgments. Theorists can avoid the moral issues by seeking psychological abilities that are common to all ways of making moral judgments, but which are not themselves ways of making moral judgments. For instance, Nichols argues that minimal mind-reading capacities as necessary to make moral judgments. Theories that offer this sort of core but refrain from insisting that one way of making moral judgments is central are subject to no moral objections.[5]

As Nichols, Hauser, Prinz, and Haidt acknowledge, the variety of ways of producing mature moral judgment gives us psychological reason to expect persisting diversity. Insofar as values are resolutely diverse, we also have moral reason to think that the mechanisms of moral judgment are plural and heterogeneous. The reason is that diverse values imply diverse

views of the good person and of the nature of morality. Psychology cannot assess these issues; this is the domain of normative theory. Without arguments demonstrating the unity of moral values, the reasonable assumption with which to approach the task of theorizing about moral judgment is a pluralist one. To begin to assess exactly what should be included in this psychological group, let's think about emotions.

2.6 Reflections on Emotions and Categorization

The role of emotions in moral judgment is a differentiating topic among Nichols, Hauser, Prinz, and Haidt. Prinz and Haidt give emotion a fundamental role in the production of moral judgment. Nichols thinks that emotions are incapable of generating moral judgments by themselves. He argues that they must be supplemented with a normative theory. Hauser agrees with Nichols about the limitations of emotions, and goes so far as to deny them any role in the generation of moral judgment. Instead, as we have seen, Hauser thinks they can merely influence the way that intuition and reason operate to produce judgments. I shall accept the assumption of Hauser and Nichols that instinct and reason can produce moral judgments. The question is whether Hauser and Nichols are correct that emotion cannot do this by itself.

The skepticism of Hauser and Nichols as to whether emotion can itself generate moral judgment stems from two simple ideas. I shall put the first in terms of judgments of wrongness. There is a distinction between something's being bad and its being wrong; not all things experienced as bad are wrong. Nichols and Hauser think that emotions are capable of delivering experiences of badness but not judgments of wrongness. Hence, either emotions must be supplemented by other mechanisms in order to produce judgments of wrongness (Nichols) or such judgments must be produced by a distinct mechanism (Hauser). For example, Nichols examines R. J. R. Blair's (1995) VIM (violence-inhibition mechanism) account of moral judgment. On this account, input of particular sorts to an affective mechanism generates feelings of aversion; these are hypothesized to account for our abilities to draw and reason about the moral/conventional distinction (Nichols 2004a, 11–12). Nichols interprets the output of the VIM as experiences of "badness" (ibid., 15), and goes on to argue that this sort of mechanism cannot deliver judgments of "wrongness" (14–16). In the following

passage, Hauser (2006, 30) agrees with Nichols: "Our emotions can't explain how we judge what is right and wrong, and, in particular, can't explain how the child navigates the path between social norms in general and moral norms in particular." Both Nichols and Hauser use moral/conventional examples to make their point. Hauser asks how emotions can teach a child that her father's anger about her eating sand indicates a conventional transgression she has made, and that his essentially identical anger about her hitting another child points to a moral transgression (ibid., 30). Nichols uses bad experiences that are not transgressions, such as falling and hurting one's knee. Although these generate aversive reactions, they do not provoke judgments that such scenarios call for punishment. However, judgments of both moral and conventional transgressions bring with them judgments that punishment is appropriate. Thus, there is an important difference between such bad experiences and whatever mechanisms realize our abilities to draw the moral/conventional distinction (Nichols 2004a, 15; see also Hauser 2006, 237–241).

The inference from cases in which emotional reactions are not tied to judgments of rightness and wrongness to the general incapacity of emotions to deliver such judgments is enthymematic. The missing premises concern a distinction drawn, in another context, by Jonathan Dancy (2004): a distinction between things that play a role and other things that either enable or disable the playing of said role, without themselves being the things that play that role. Dancy's topic is moral reasons. For example, that something causes pain can be in one context a reason not to perform an action and in another context no reason at all either in support of or in opposition to the performance of some action. There are two possible differences between these situations. It might be that pain is enabled to play this role in the first context but not in the second. For instance, the pain might be intentionally caused in the first but not in the second. In Dancy's terms, we might suppose that the intentional status of the pain enables the pain to function as a reason without itself being a part of the reason, so that, when asked why we should not perform the action in question, we can fully and truly answer "Because it would hurt." Second, it might be the case that pain itself usually suffices to function as a reason of this sort, but that in the second case there is something that disables its capacity to do so. Imagine that, in the second case, the action in question would save a thousand lives. The stakes are so high that the fact that the

action would cause pain is now irrelevant—it provides no reason against the action in question, in contrast with the first situation.[6]

Here is how these distinctions apply to the present topic: Following Dancy, let's call properties that enable something else to play a role "enablers" and properties that prohibit other things from playing a role "defeaters." We cannot justifiably conclude that the failure of emotions to generate judgments of rightness or wrongness in specific situations indicates that they can never play this role without examining possible enablers and defeaters. That these considerations apply to psychological mechanisms is clear enough. For instance, imagine that I duck in response to a shadow that appears near my head. The shadow is the object of my perception and the trigger of my response. My visual system enables this response, but it is not itself a part of the representational content to which I am responding. In the cases offered about moral judgments and emotions, perhaps emotions are prevented—i.e., defeated—from generating proper moral judgments by features of these particular situations. Or—what I find more likely—perhaps there are features of some situations that enable emotions to generate moral judgments without themselves being parts of the mechanisms that generate moral judgments. Both Hauser and Nichols move from cases in which emotions fail to generate moral judgments to the conclusion that something else must play a role in producing them. However, the conceptual possibilities are subtler than this, and thus the inference made by Hauser and Nichols is unduly hasty.

Though it is important to respect the difference between badness and wrongness, this line of thought about what emotions can and cannot do deserves scrutiny. Other than the ideas just examined, neither Nichols nor Hauser offers any psychological evidence for the judgment that emotions can deliver verdicts of badness but not wrongness. They treat it as relatively obvious, but this might be a failure of imagination. It certainly seems possible that one might deny that one shares the intuitions that Nichols and Hauser have and encourage them to rethink emotions. Perhaps, phenomenologically, some people genuinely feel some actions to be wrong, not merely bad. More subtly, Prinz ties judgments of wrongness to a particular range of emotions. When we feel something to be wrong, the emotion stems from the distinctively moral disapprobation range. When we feel something to be bad, however, we are experiencing emotions from some other range.

Besides cases and the first-person experience of emotions, skepticism about the role of emotions in moral judgment seems to be driven by ideas about categorization. Nichols supplements emotions with a normative theory that groups actions into permissible and impermissible. Hauser agrees with the spirit of this suggestion, but instead argues that the necessary ingredient is an instinctive mechanism that analyzes actions in terms of their causes and consequences. Both think that emotions cannot perform the relevant kinds of categorization themselves.

There are empirical grounds to question this assumption, although not to reject it outright. Consider so-called mirror neurons, which have been discovered in monkeys and in humans. (Prinz discusses mirror neurons briefly on page 229 of his 2004 book; see also Rizzolatti et al. 1996, Keysers et al. 2003; Gallese et al. 2004, and Rizzolatti and Craighero 2004.) These neurons are activated both when a monkey is observing another monkey perform an action of a particular kind and when a monkey is itself performing an action of the same kind. The evidence also suggests that the same brain structures are responsible for both production and recognition of emotions. There is a general lesson here: we cannot move directly from *conceptually* distinct ideas, such as action-production and action-categorization, to the positing of *psychologically* distinct mechanisms or structures. Conceptually distinct functions, it turns out, can be realized by the same mechanism. To apply this to the present topic: the move from the *conceptual* distinction between badness and wrongness to a *psychological* distinction between emotions capable of detecting badness only and other mechanisms that deliver verdicts of wrongness must be handled carefully.

This, of course, leaves all the important questions unanswered. I shall return both to mirror neurons and to the mechanisms by which we make judgments of wrongness, not mere badness, in a later section on the mechanisms of moral judgment. To get there, I must first examine some of the details of Haidt's social-intuition model of moral judgment.

2.7 Externalization and Process in the Social-Intuition Model of Moral Judgment

Although Haidt calls his position a "social intuition" model of moral judgment, he shares at least as much with Hauser as he does with social cognitivists such as Elliott Turiel (1983) and Judith Smetana (2006).[7] Like

Hauser, Haidt uses an analogy between language and moral judgment: "The social intuitionist model . . . proposes that morality, like language, is a major evolutionary adaptation for an intensely social species, built into multiple regions of the brain and body, that is better described as emergent than as learned yet that requires input and shaping from a particular culture." (2001, 826) There is, however, a *prima facie* tension in Haidt's position between innate and social contributions to moral judgment. Insofar as this tension stems from one way of developing the linguistic analogy, it might be shared by Hauser's position. In this section, I shall focus only on Haidt's work.

Haidt uses the metaphor of *externalization* in describing the process by which an innate contribution is made to moral judgment (2001, 826–827). He is clear that he thinks that this is only one part of the story about moral judgment (ibid., 826). However, this contention requires attention. Haidt applies the notion of "externalization" to moral intuitions. Recall that for Haidt intuitions include but are not limited to emotions. Presumably what holds for emotions in general holds for the intuitive sources of moral judgment more particularly. Emotions cannot literally be externalized. If I am feeling happy, I cannot literally bring this out into the public domain. Instead, I can, in a sense, externalize my happiness via expression or display in various kinds of utterance or behavior.[8] Moreover, the idea of "externalization" suggests that something complete and sufficient in itself is brought out into the open via the expression in question. But this is not what Haidt intends for moral judgment. Recall that he emphasizes fluid, taking-place-through-time processes of moral judgment including social influences. If this is apt, we have no reason to see moral judgment as a process of bringing out into the open something already formed but located within the person. Instead, there is some sort of contribution made by a person, but this is subject to modification through social processes. Overall there seems to be a conflict between the way Haidt attributes an innate aspect to the origin of moral judgments and the way he casts moral judgment as involving interpersonal processes. To put the problem differently: If people can externalize solely that which is complete within them, then there is no room for social influence on the products of such externalization. If moral judgment is *mostly* a process of externalization, then there is theoretical pressure against seeing moral judgment as *typically* a social, interpersonal process.

This conflict can be alleviated by rejecting the way in which Haidt presents the overall process of moral judgment and adopting another. Haidt draws a sharp contrast between views that present moral judgment as "a single act that occurs in a single person's mind" (2001, 828) and the social intuitionist portrayal of judgment as a temporally extended interpersonal process. However, the portrayal of judgment as an extended process is inconsistent with Haidt's way of describing the innate aspect of moral judgment. This conflict is resolved if we see the process in question as constituted by a series of different sorts of moral judgments. On this view, the first stage might well be a matter of expressing the emotional intuitions an individual has by virtue of biology. Subsequent stages are a matter of using this judgment, and/or the information that gave rise to it, as fodder for *further* judgments, including both judgments made solely by the individual who made the original judgment and judgments made by that individual in collaboration with others. The interpersonal processes in question include, but probably are not limited to, the explicit use of moral reasoning. Such a view of moral judgment allows for a substantial role for emotion in the generation of moral judgments, for a substantial sense of in which moral judgment can be a matter of "externalization," and for a substantial and simple sense in which moral judgments are produced by interpersonal processes.

On this view, moral judgment can be both a process of externalizing something inner and a process of collaborating with others. However, there is a theoretical price to be paid for construing the process of moral judgment as constituted by a series of judgments. Insofar as such a process does not require a single kind of beginning, this sort of process is performatively disunified. That is, giving this sort of process an important place in one's account of moral judgment amounts to portraying the mechanisms of moral judgment as fundamentally plural in kind: some are emotional, some are reasoning, some are intuitive in some third sense. There is no single psychological capacity that constitutes the performative core of moral judgment on such a view. I think that this is the correct way to see moral judgment.

2.8 The Mechanisms of Moral Judgment

If I am correct, then we have good reason to reject unity theories of moral judgment and instead pursue theories that portray it as performatively

disunified.[9] However, much can be learned and retained from the work of unity theorists even if we do this. Besides illuminating the roles of our emotional, reasoning, and instinctual capacities in moral judgment, the work of Hauser, Prinz, Haidt, and Nichols helps us to identify two desiderata for theories of moral judgment: Such theories should include (1) an innate biological component or components to account for the early, regular development of moral-judgment capacities revealed in such studies as those on the moral/conventional distinction and (2) a social component or components to account for cultural variety in mature processes of moral judgment. In principle, these desiderata can be satisfied either by a single aspect of the overall account of moral judgment or, which is more likely, by distinct mechanisms.

In the remainder of the theoretical portion of this chapter (that is, through section 2.10), I will briefly sketch a hypothesis for a mechanism of moral judgment to be added to those provided by Nichols, Prinz, Haidt, and Hauser. This mechanism will be described from an explicitly externalist point of view. I will address desiderata 1 and 2, emphasizing social processes.

Nichols and Hauser both think that emotions could not themselves be the origin of moral judgment because, although they could deliver verdicts of badness, they could not evaluate an action or a state of affairs as *wrong*. Let's put aside the fact that Nichols and Hauser do not argue for this notion and grant it for the sake of argument. Even having done so, we need not follow Hauser and Nichols in supplementing emotions with such constitutively distinct mechanisms as an instinctual action analyzing capacity (Hauser) or a process for incorporating normative theory (Nichols).

In his embodied appraisal account of the nature of emotion, Prinz (2004, 234–236) distinguishes two sorts of pathway that constitute emotional processing.

First, there are *initiation* pathways. These may be thought of as the input routes to the emotional module. Their general job is to receive input from a variety of sources, and then to prepare this input in a manner appropriate to the remainder of the emotional processing. As an example, Prinz discusses the role of the amygdala in the processing of fear, disgust, and sadness: "The amygdala receives inputs from a variety of different brain regions and initiates a pattern of bodily outputs, which then give rise to these emotions." (2004, 234)

An important feature of the initiation pathway is what Prinz calls *calibration files*. Calibration files are sets of representations linked to particular bodily responses. Prinz holds that such files allow us to modify emotions via judgments. The establishment of new calibration files allows us to modify emotions (more specifically, embodied appraisals) to apply to things other than those to which they evolved to apply (2004, 99–100).

Second, there are *emotion response* pathways. Crucially, Prinz holds that this is where the actual emotions are realized. Accordingly, emotional processing has the structure illustrated in figure 2.1. Strictly speaking, on this view the initiation pathways are constitutively distinct from emotions themselves. However, they, including their calibration files, are a part of the production of all emotions (Prinz 2004, 101–102). If we apply this model of emotional processing to the line of thought offered by Hauser and Nichols, then evaluations of *wrongness* have to be brought about by influences on the initiation pathways, not on the response pathways. If emotional-response pathways are to be sensitive to wrongness, they have to be calibrated for this.

The strategy suggested by the combination of the work of Hauser, Nichols, and Prinz is individualistic, in that the supplement to emotions proper is some distinct in-the-head psychological mechanism for the production of judgments of wrongness. However, externalist models of cognitive processing suggest alternative ways of supplementing the emotional-response pathways. Instead of input being provided by a mechanism within that individual that has the function of producing assessments of wrongness, the initiation pathway could in principle be constituted by a mechanism that tracks external resources. In this case, the relevant external resources are, first and foremost, the judgments made by

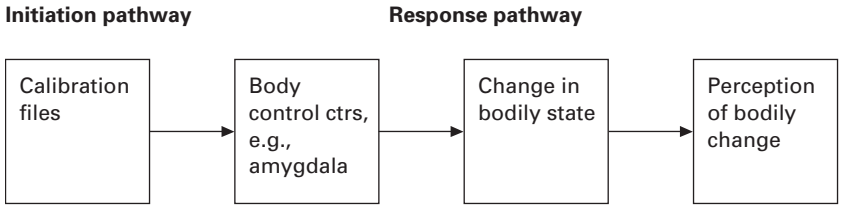


Figure 2.1
The structure of emotional processing. Adapted from Prinz 2004, 235.

other people, as displayed in their public thought, talk, facial expressions, and behavior.

An obvious suggestion for the neural basis of this ability is mirror neurons. One of the striking features of mirror neurons is how they seem to connect us to others. Increasingly researchers point to these structures as the foundation of social cognition. (See, e.g., Gallese et al. 2004; de Vignemont and Singer 2006; Singer 2006.) These neurons bear the name "mirror" because of the way they reflect what is seen in others in portions of the observer's brain responsible for the production of the effect in the observer. Mirror neurons have been found for the recognition and production of emotions and for the recognition and production of actions. My hypothesis is that mirror neurons may be involved in the recognition and the production of moral judgments. Specifically, they may be a neural bridge for calibrating one's moral judgment with those of the people around one.

So far I have focused on emotions. This is due to the attention given to emotions by the theorists discussed in this chapter, and to the work on mirror neurons in emotional processing. In the schema of chapter 1, that makes the current externalist hypothesis a *shallow* one: a psychological function attributed to individuals regardless of environmental integration has been reinterpreted in an explicitly externalist way. But the present line of thought applies more widely. *Any* mechanism (not only a mechanism tied to emotional processing) that is simultaneously capable of tracking the judgments of others and producing concordant categorizations of actions and events is a candidate for an externalist mechanism of moral judgment. This deeply externalist hypothesis is *extremely* speculative but nonetheless interesting.

Of the theorists examined in this chapter, the one who has the most in common with the present suggestion is Haidt. Haidt gives emotions a direct role in the intuitive basis of moral judgment. He also describes moral judgment as an interpersonal process. However, there are differences between Haidt's position and the current suggestion. For one, Haidt gives other people a role in the production of moral judgment primarily through interpersonal reasoning processes. In contrast, the present position proposes emotional and other non-reasoning processes by which the judgments of others can have a constitutive role in the production of an agent's moral judgments. Second, and relatedly, Haidt casts his position in terms of individualistically construed processes. The present position points to a

role for hybrid processes in moral judgment—processes that involve deep functional and causal integration of an individual with features of the environment.

2.9 Some Thoughts about Embeddedness

I have argued at medium length that we should take a pluralist view of the mechanisms of moral judgment, and I have suggested at much shorter length that some of these mechanisms might be individual-world hybrids. What of the "embeddedness" I mentioned at the outset? How is this an account of plural, hybrid, *embedded* moral judgment?

Consider the topics that have arisen in this chapter: emotion, reasoning, the ways in which moral-judgment processes can involve other people. Consider also the importance of carefully handling the inference from conceptually distinct ideas to psychologically distinct mechanisms. My suspicion is that moral judgment, in many if not all of its guises, is carried out simultaneously and by the same mechanisms by which we attribute responsibility, produce action, and reason. That is, I suspect that our capacities for moral judgment may be largely embedded in other aspects of our moral psychology. Though this may accentuate the importance of attending to moral judgment in one sense (it's everywhere), in another sense it downplays its importance. If moral judgment is embedded in other capacities, then psychological theories should not be designed around free-standing moral-judgment capacities. If this suspicion turns out to be correct, then the methodological orientation of recent psychological work on moral judgment is mistaken, and the present emphasis on other psychological processes as of equal importance is vindicated.[10]

Let's draw some distinctions. Embeddedness is a relation between psychological functions: the present topic is ways in which one psychological function P can be embedded in another psychological function R. P and R are conceptually distinct psychological functions, but embeddedness is an issue of realization (implementation, the means of performance). A useful way to think about embeddedness is in terms of systems. Here is a first pass: P is psychologically embedded in R when the systems by which P is realized are the same as those by which R is realized. Whether any psychological functions share systems in this way is an empirical issue, and hence cannot be decided by merely *a priori* means.

To allow *a posteriori* evaluation of hypotheses about embeddedness, we can make our understanding of this notion more fine-grained using the notion of a system:

(A) P is *individualistically* embedded in R just in the case that P shares with R portions of a realizing system that are located with the physical bounds of an individual.

(B) P is *environmentally* embedded in R just in the case that P shares with R portions of a realizing system that are located beyond the physical bounds of the individual who has P and R.

Both (A) and (B) deliberately make reference to "portions of" realizing systems to acknowledge the possibility of degrees of embeddedness. We can use this idea for a second distinction:

(C) P is *completely* embedded in R just in the case that (i) P is at least individualistically embedded in R and (ii) no aspects of P are not embedded in R.

(D) P is *partially* embedded in R just in the case that (i) P is to some degree embedded in R, either individualistically or environmentally and (ii) P is to some degree not embedded in R, either individualistically or environmentally.

(C) emphasizes the individualistic aspect of psychological functions because this is shared by both narrow and wide hypotheses about psychological functioning. For a psychological state to be attributed to an individual, at least something, and maybe everything, must be going on within the bounds of the individual that realizes that psychological state. By combining (A) & (B) with (C) & (D), we generate an array of conceptual possibilities:

(NC) *Narrowly complete embeddedness* P is individualistically embedded in R such that there no individual-attributable aspects of P are not embedded in R.

(NP) *Narrowly partial embeddedness* Some but not all of P's individual-attributable aspects are embedded in R.

(WC) *Widely complete embeddedness* P is environmentally embedded in R such that no environmental aspects of P are not embedded in R.

(WP) *Widely partial embeddedness* Some but not all of P's environmental aspects are embedded in R.

The four possibilities can also combine. For example, besides NC alone, in principle two psychological functions can be related in any of the following ways:

NC & WC

NC & WP

NP & WC

NP & WP.

So much for conceptual distinctions. Do we have any reason to think that moral judgment really is embedded in other psychological abilities in any sense? Let's begin with the hybrid mechanism of moral judgment that was sketched in the preceding section. I suggested that the combination of emotions and the capacity to participate in interpersonal reasoning systems, and in particular to track the judgments of others, could constitute a mechanism for moral judgment. In terms of the other psychological capacities examined in this book, this hypothesis embeds a mechanism for moral judgment in moral reasoning. Although it is not clear exactly how such embeddedness works, in this case it involves at least environmental embeddedness: external resources available via interpersonal reasoning capacities provide the supplement needed for emotions to produce specific judgments of, e.g., wrongness rather than mere badness.

The discussion of moral reasoning in the next chapter suggests deeper embedding. Many influential ways of doing psychological research, including those in the tradition of studying the moral/conventional distinction, are aimed at judgment. Moral/conventional-distinction tests are explicitly aimed at moral judgment. However, insofar as these tests require the use of information about morality, they are first and foremost tests of moral reasoning. They shed light on moral judgment by getting subjects to deploy their abilities at the conscious, intentional transformation of information about moral issues. The more intimately moral judgment is connected to such reasoning, the more light these tests can shed on moral judgment. Embedding is a particularly intimate way in which moral judgment can be related to moral reasoning. Once we give up on the search for "core" moral

judgment, and once we reflect on the differences between conceptual and psychological distinctness, we have reason to take seriously as a hypothesis that one way moral judgment is performed is via both narrowly and widely complete embedding in moral reasoning. That is, our capacities for moral reasoning are inherently capacities for moral judgment.

Finally, let's reflect on the significance of automaticity. Both Haidt and Hauser emphasize the automaticity of moral judgment as counting against moral reasoning's being its source. Let us follow them in this and think about other aspects of our moral psychology. Moral judgment is not the only thing we do automatically. It seems to me that, some of the time, we attribute responsibility and produce actions automatically. For instance, some ways of deploying what P. F. Strawson famously called the "reactive attitudes," such as resentment, are much like automatic reflex responses. The same holds for action. Some of the behavior for which we attribute moral responsibility seems to be performed automatically given the situations in which agents find themselves. One person succumbs to temptation without thought and shoplifts; another automatically helps somebody who has slipped on ice. Again, once we give up on the idea of "core" moral judgment, and once we attend to the differences between conceptual and psychological distinctness, such automaticity provides *prima facie* support for the framing of hypotheses that embed some ways of performing moral judgments *completely*, both narrowly and widely, in our capacities for producing actions and attributing responsibility. That is, these capacities amount to ways in which we perform moral judgment.

2.10 Theoretical Conclusion

The prevailing "judgment-first" methodology of moral psychology smacks of Rylean intellectualism. Gilbert Ryle (1949) criticized accounts of the mind, and especially of the production of behavior, that emphasized the psychological precedence of explicit thought. I suspect something similar is apt for the prevailing trend of approaching moral psychology via moral judgment. Why should we be so certain that action, moral reasoning, or attributions of moral responsibility are guided by psychologically distinct mechanisms of moral judgment?

Reflect on "unity" approaches to moral judgment and the individualism/externalism theme of this book. Theories of moral psychology tend to

address moral judgment as the primary moral-psychological constituent. They also tend to explain moral judgment in terms of a single psychological "core" capacity. Although such theories address social influences on moral judgment, the prevailing approach is to explain moral judgment from the inside out. This approach has implications for other aspects of moral psychology. Perhaps the intellectualist judgment-first approach does not derive explicit support from the tendency to devise unified accounts of moral judgment, but they are, shall we say, mutually resonating. If my arguments in this chapter against unity accounts of moral judgment are cogent, the judgment-first approach begins to look suspiciously lonely. The rejection of a unity structure for theories of moral judgment brings with it pluralism about the mechanisms of moral judgment. This, to my theoretical ear, resonates more with externalist and embedded views of moral judgment and our overall moral psychology than with intellectualism and individualism. Once we reject the unity structure, the variety of forms in which we find mature moral judgment looks fundamental rather than derivative. This invites theories with diverse mechanisms, some of which are widely realized. My opinion is that such spread of moral judgment makes it look less special and more like merely an important aspect of other sides of our moral psychology. Such suspicions, however, can only be assessed over the long term, as data are collected and theories are tested.

2.11 Application: Moral Dilemmas

If I am right about the plural, embedded, and hybrid aspects of moral judgment, it is reasonable to think that some phenomena studied under the auspices of moral judgment are actually produced in more complex ways that involve both external resources and other psychological capacities. In the remainder of the chapter, I will develop such a case in connection with studies of judgments produced about moral dilemmas. Specifically, I shall discuss extant accounts of this topic, generate a rival hypothesis on the basis of the discussion in this chapter, and sketch how it might be assessed empirically. This exercise will show how externalist hypotheses can make a difference to the practice of both empirical and philosophical psychology.

In the last few decades, moral philosophers have noted an asymmetrical pattern in responses to hypothetical moral dilemmas. The so-called trolley

cases ask people to imagine responding to a runaway train. Here are two brief versions of this scenario:

(A) There is a runaway train. If left alone, it will hit and kill five people. You can divert it to a different track. This will save the lives of the five people, but kill a sixth. These are your only options. Is this permissible?

(B) There is a runaway train. If left alone, it will hit and kill five people. You know that pushing the very large person beside you onto the track will stop the train; sacrificing yourself will not succeed in doing this. This will save the five lives, but kill the person you push. These are your only options. Is this permissible?

The *a priori* conviction of philosophers has been that generally people think that diverting the train is permissible but that pushing the single person is not.[11] In the last few years, the trolley cases have been addressed empirically. The empirical studies have confirmed that the asymmetry in fact characterizes how people respond to the dilemmas. Given that the life/death ratio is the same in both cases, the difference in response is puzzling. Why is it the case that people answer as they do? I shall examine two important hypotheses that have emerged to account for this asymmetry. Joshua Greene and colleagues have provided an emotion-based explanation of the asymmetry (Greene et al. 2001; Greene and Haidt 2002; Greene et al. 2004). On this account, emotions are engaged in the pushing case that are not engaged in the diverting case, and this emotional difference gives rise to the asymmetry. More recently, Shaun Nichols and Ron Mallon (2006) have offered a rule-based explanation. On their view, the asymmetry is due to the cognitive importance of rules: there is a general rule against direct killing of the sort found in the pushing scenario, but no rule against the inadvertent causing of death that comes with the diverting of the train, so subjects judge that the latter and not the former is permissible.

I shall argue for the following three theses:

Both views are problematic.

Each of the two views contains a grain of insight into moral dilemmas.

A third hypothesis, the *conformity* explanation, is at least as plausible as either of the competing explanations.

In a nutshell, the conformity explanation holds that the asymmetry is due to the cognitive importance of conforming one's view of the world, as it is revealed through one's judgments and behavior, to that of other people. At base, this is an emotion-based hypothesis, so it shares more with the position of Greene et al. than with the rule-based position of Nichols and Mallon. Nevertheless, it is best seen as a distinct third option. Before getting to it, I will examine the extant emotion-based and rule-based explanations.

2.12 Emotion-Based and Rule-Based Explanations of the Asymmetry

Greene et al. (2001, 2106) argue that certain emotions are engaged in the pushing case that are not engaged in the diverting case because the pushing case involves personal interaction with the person who will be killed, whereas the diverting case does not: "The thought of pushing someone to his death, we propose, is more emotionally salient than the thought of hitting a switch that will cause a trolley to produce similar consequences, and it is this emotional response that accounts for people's tendency to treat these cases differently." The empirical basis of this proposal is functional magnetic resonance imaging (fMRI) of neural activity in people who are considering the trolley cases. Greene et al. report more activity in Brodmann's areas 9 and 10 (medial frontal gyrus), 31 (posterior cingulated gyrus), and 39 (angular gyrus, bilateral) when people think about the pushing case than when they think about the diverting case. They cite studies associating these areas of the brain with emotional processing (ibid., 2107). Hence their hypothesis that increased emotional engagement accounts for the asymmetrical pattern of response to the trolley cases.

Nichols and Mallon present two lines of objection to this account. First, the increased emotional engagement is supposed to derive from the personal interaction characteristic found in the pushing case but not in the diverting case. However, it is not hard to think of real-world cases that are characterized by such personal engagement, and even by high emotional arousal, in which the infliction of harm or death is seen as permissible. Nichols and Mallon offer as examples circumcision of male infants, certain acts of war and self-defense, and certain acts of punishment (2006, 532). Here we get emotionally charged personal interaction coupled with

judgments of moral permissibility, which directly counters the conjunction of personal interaction and emotional engagement with moral impermissibility that is at the heart of the emotion-based explanation of Greene and colleagues.

The second line of criticism of emotion-based explanations presented by Nichols and Mallon stems from new studies. They hypothesize that the differentiating cognitive factor in the asymmetry derives from the role of rules in practical thinking, not from emotions (2006, 533–434). If they are correct, then this asymmetry should be found in "*impersonal* scenarios with minimized emotional content" (ibid., 534). To test this, Nichols and Mallon devised two new scenarios not about life and death but about breaking teacups. Otherwise these cases are analogous to the trolley cases. In the case analogous to the trolley case about diverting the train, a child's mother explicitly forbids him from breaking teacups that are on the kitchen counter. Billy (the child) later sets up his model railroad, then becomes distracted by a snack. He returns to discover that his sister Susie has placed teacups on the tracks. If the train stays on its present course, it will break five cups. Billy has only enough time to divert the train with a lever, which will result in breaking a solitary sixth cup. He diverts the train (ibid., 534–535). In the case analogous to the trolley case involving pushing someone onto the tracks, again a child's mother explicitly forbids anyone to break teacups. Susie, later discovers her younger brother playing with the teacups and a toy truck. The truck is about to break five teacups. Susie is next to the counter on which the other teacups sit. The only way to save the cups is to throw a solitary sixth cup at the truck, thereby changing its course. Susie can throw well and knows that she will succeed if she throws the cup. She throws the cup, breaking it but saving the others (ibid., 535). After each of the aforementioned cases, subjects were asked whether Billy or Susie broke the mother's rule, and whether this was, all things considered, acceptable. In the first version of the study, all subjects answered that the rule was broken in the teacup-pushing analogy; however, for every two subjects who said that the rule was broken in the third case, one said that is wasn't (ibid., 535–536). In the second, between-subjects version of the study, "96% of the participants said that a rule was broken in the [pushing] case, but only 44% said that a rule was broken in the [diverting] case" (ibid., 536). Overall, these experiments seem to provide a correlation between recognition of rules and judgments of impermissibility.

The broader empirical background of this appeal to rules in moral judgment is the tradition of moral/conventional distinction. Nichols and Mallon (ibid., 533) cite a 1987 study by Elliot Turiel and colleagues as showing that judgments about conventional violations depend, in part, on knowledge of local rules. Along the same lines, R. J. R. Blair and colleagues (1995, 18) have found references to rules in explanations of judgments about moral and conventional transgressions.

This recent history bears the hallmarks of intellectual progress: a hypothesis is formulated and tested, then a different hypothesis is formulated and evidence is advanced that supports it rather than the original one. Nichols and Mallon present empirical results that seem to count against the role of emotion in thought about dilemmas. Contrary to appearances, I shall argue that we are not substantially closer to understanding the psychological roots of this phenomenon than we were before the development of these emotion-based and rule-based hypotheses. One reason for this is internal to the particular studies that have been performed. However, a subtler reason stems from the relation between these studies and other developments in moral psychology more generally. To see this, let's consider the explanatory task at hand.

2.13 Three Explanatory Desiderata

Here are three desiderata of adequate explanations of the asymmetrical pattern of response to hypothetical moral dilemmas:

Such explanations should address causal mechanisms by which the responses in these experiments are produced.

The causal mechanisms must account for all of the relevant data.

The explanation must be sensitive to the difference between the psychological origins of moral judgment and influences on processes by which such judgments are developed and publicly performed.

The first two of these desiderata are uncontroversial, but the third needs defense.

Here are two ways to think about the importance of distinguishing between the psychological origins of moral judgments and the influences on the development and public performance of these judgments.

First, consider the explanatory target and the body of evidence in question. The extant explanations try to address a psychological competence, i.e., the psychological capacities that generally make it possible to make judgments of this sort. But the explanandum is not merely a type of judgment; it is a judgment made explicit in speech or writing. It is reasonable to think that, besides our general competence with in making moral judgments, such performances draw on other, distinct psychological resources. As Hauser et al. note about linguistic performances as opposed to competences: "What this individual chooses to say is a matter of her *performance* that will be influenced by whether she is tired, happy, in a fight with her lover, or addressing a stadium-filled audience." (2008, 111) The list of influences could be extended significantly. We should take seriously the idea that these performance processes affect the content of the expressed judgments. It is unduly naive to think of such judgments as necessarily or even typically being passed on without transformation from our initial judgments to our public performances.

Second, and more subtly, once we reflect on the processes involved in getting from a psychological capacity to a public performance, we are in a good position to take seriously the idea that the process of moral judgment itself, before being publicly performed, is temporally extended. This opens up the possibility that there are important differences between the psychological origins of moral judgment and subsequent psychological processes of moral-judgment formation. As we have seen, some of the leading accounts of the psychology of moral judgment make exactly this distinction. I have already examined the details of Haidt's position on exactly this point. Hauser's account is another example. Recall that Hauser draws an analogy between moral judgment and our linguistic capacities, and on this basis hypothesizes that we have a "moral instinct." (See, e.g., Hauser 2006, 32–42.) The moral instinct is the origin of our moral judgments. However, Hauser thinks that the judgments produced by the moral instinct are subject to modification or even opposition from reason and emotion. Overall, the picture of moral judgment provided by Hauser has places both for initial psychological sources of moral judgment and for subsequent processes of forming moral judgments.

For present purposes we need not choose between Hauser and Haidt, or between these and other accounts of the psychology of moral judgment. All that is required is acknowledgment that the field is characterized by a

distinction between the origins of moral judgment and subsequent, distinct psychological processes of moral-judgment formation and of the public performance of moral judgments. It is this general distinction, not the details of any particular account of moral judgment, that is important for assessing the two accounts of the asymmetrical pattern of responses to hypothetical moral dilemmas.

2.14 Assessment of Extant Emotion-Based and Rule-Based Explanations

Nichols and Mallon are correct about the position of Greene et al.: it fails to account for all of the relevant data. Specifically, it fails to account for the emotionally neutral yet asymmetrical pattern of results presented by Nichols and Mallon. Neither does this position distinguish between the origins of moral judgments and subsequent processes of judgment formation and their public performance. The rule-based explanation offered by Nichols and Mallon is equally problematic. Specifically, it fails by the standards of the first and third explanatory desiderata offered in the previous section.

Let's begin with the requirement that an explanation illuminate the causal mechanisms responsible for the asymmetrical pattern. Nichols and Mallon have as a part of their foundation the tradition of research into the moral/conventional distinction. However, this tradition allows for normative judgments in the absence of rules. This has been a feature of such research for decades. In a 1981 study that is typical of this tradition of research, Judith Smetana asked children whether a depicted event would be acceptable "if there was no rule about it" (1334). In a 1995 study, Blair used two questions that were virtually the same: "If there were no rules about people doing (the transgression), would it still be a bad thing to do?" "If there were no rules about people doing (the transgression), would it still be a good thing to do?" (190)

In an analysis of justifications of judgments offered by normal people and by psychopaths, Blair found that normal people appealed to rules (in the broad sense) *at most* one-third of the time. Psychopaths cited rules roughly 40–50 percent of the time (ibid., 18). The import is that sensitivity to things *other than rules*, such as others' welfare, is part of what the moral/conventional distinction tradition suggests is responsible for the sorts of evaluative judgments that people make.[12]

In view of this background, tests must be carefully designed in order to reveal the causal efficacy of rule-cognition in judgment. Nichols and Mallon's teacup cases were not adequately designed to reveal this sort of role for rules. To simplify a bit: In their experiments, subjects were asked two questions. The first was whether the action was permissible; the second was whether Billy or Susie broke a rule (2006, 535, 538). The answers to these questions were compared, revealing correlation between judgments of impermissibility and assessments of when a rule was broken. However, for the purposes of Nichols and Mallon it was dubious to ask explicitly whether a rule had been broken in these cases. That was a leading question. Positive answers revealed agreement of the subject with an after-the-fact interpretation of the cases provided by the experimenters. They did not necessarily reveal that sensitivity to rules was a causally efficacious factor in the production of the judgment of impermissibility.

This issue is a particular instantiation of one that has been repeatedly raised in psychology. Whenever psychologists study the responses of people to particular prompts, it is relevant to ask just whose characterization of the prompts is the important one: that of the experimenters, or that of the subjects. The general consensus is that, in order to shed as much light as possible on the psychological processes responsible for judgment and behavior, it is the subjects' own characterization of the situation that is the important one. For specific discussion of this point, in the "person-situation" debate, both philosophers (e.g., C. Miller 2003, Sreenivasan 2002, Kupperman 2001) and psychologists (e.g., Mischel 1999, Mischel and Shoda 1995; for discussion, see Doris 2002, 76–85) make this claim.[13] In official terminology, descriptions provided by experimenters are "objective" (Ross and Nisbett 1991, 11; Sreenivasan 2002, 50) or "nominal" (Doris 2002, 76). Descriptions provided by subjects are "subjective" or "psychological."

Here is how this applies to the present issue. By asking whether a rule was broken, Nichols and Mallon risked leading their subjects by substituting a nominal interpretation of the situation for a subjective/psychological one. Arguably such a method cannot be trusted to provide reliable information about the role of the content of the interpretation—in this case, rules—in the production of the subjects' judgments. A better experimental design would be one in which subjects were asked to provide their own

explanation of their judgments, without any leading suggestions. Blair did this in his study of psychopaths and the moral/conventional distinction. After being asked about the permissibility and seriousness of an act, subjects were asked "Why was it bad for X to do [the transgression]?" (Blair 1995, 15) As we have seen, although rules show up in the psychological answers to this question, they do so at most one-third of the time for normal subjects. Blair reports that the welfare of the other was most commonly used to justify moral judgments (ibid., 18). On the assumption that answers to questions of this sort can reveal information about the causes of judgments (an assumption to which Nichols and Mallon are committed by their experimental design), the findings of Blair complicate their hypothesis. On the basis of Blair's information, one would predict, looking at statistical prominence alone, that it would be something to do with response to the other's welfare that would generate the asymmetry in the trolley cases. In order to assess the role of rules against this empirical background, tests must be very carefully and precisely designed. Not only must they show a role for rules; they must distinguish rules from other elements as the cause of the asymmetries. The experimental set-up of Nichols and Mallon fails to do this. The implication is that there is less empirical support for the role of rule-cognition in the generation of the asymmetries than Nichols and Mallon think.

Now let us turn to the remaining issue: distinguishing between the origins of moral judgment and subsequent processes of judgment formation and public performance. This distinction presents the following possibilities: the asymmetrical pattern of responses to hypothetical moral dilemmas might be due to the psychological starting point of moral judgment, or it might be due to downstream psychological processes of the development and the expression of moral judgments, or both. The extant positions neglect this distinction and this array of possibilities. Both of them exemplify the first possibility. One might think that recognition of this distinction merely shows that such hypotheses require supplemental refining. To show the importance of this distinction, and hence to show that the problem faced by extant positions is more serious than a mere unfilled gap, I shall make a first stab at explaining the asymmetry in terms of subsequent influences on moral judgment and its performance rather than in terms of its origins. I call it *the conformity hypothesis*.

2.15 The Conformity Hypothesis

To get to the conformity hypothesis, let's give the emotion-based explanation a second thought. Although they originally put their position in terms of *emotional* engagement and *personal* interaction, in a later paper (2004) Greene and colleagues repeatedly use the term 'social-emotional'. They do not see this as a significant change. However, I think that the move away from the personal and (especially) the emotionally charged to a position that focuses on social interaction is significant. For one thing, it connects this topic with a large body of empirical studies of social interaction. For another thing, it promises to improve upon the narrow scope of the emotion-based position of Greene et al., thereby avoiding the problems identified by Nichols and Mallon. I shall now address these issues in order.

Social psychologists have spent decades studying processes of conformity. Some of this research, such as studies about conforming to the desires or instructions of others, overlaps with the person-situation debate. Stanley Milgram's infamous studies on obedience (1963) are the best known. Milgram solicited participation in learning studies, but this was a set-up. Subjects were given the role of teacher, while confederates of the experimenters played the roles of learner and of study administrator. The job of the teacher was to ask questions and administer electric shocks in response to incorrect answers. The shocks ascended in severity in 15-volt increments clearly labeled with serious warnings. When subjects hesitated in administering shocks, the administrator-confederate politely recited a list of instructions to continue. Milgram found that features of experimental situations that were by normal standards non-coercive seemed to lead ordinary people to administer what they thought were lethal levels of electrical shocks to other ordinary people. More precisely, about two-thirds of subjects administered shocks all the way up to the final level, and many of the other subjects administered shocks up to very high levels.

Other studies have addressed conformity of judgment rather than behavior. The most famous studies of conformity of judgment are those of Solomon Asch. In one simple experimental set-up, Asch had groups of seven to nine persons judge which of three lines in various groups matched a standard line. All the judges except one were confederates of the experimenters. In the first three judgment cases, the confederates gave the obviously correct answer. But in the fourth case, the confederates, answering

before the actual subject, all chose an obviously incorrect answer—a 6-inch line rather than an 18-inch line. In this set-up, 50–80 percent of subjects conformed their judgments to that of the group rather than going with the clear and simple evidence that was before their eyes (Asch 1951, 1952, 1955, 1956; Ross and Nisbett 1991, 30–32).

John Sabini and Maury Silver (2005) have offered an explanation of the results of the conformity experiments that focuses on social cognition, i.e., the psychological significance to an agent of the views of other agents. Sabini and Silver argue that the effects on behavior are brought about through the nuances of social interaction. Embarrassment and confusion brought on by the prospect of behaving in ways that show that one sees the world in a way different from others in the same situation are the emotional responses to social pressures offered by Sabini and Silver (2005, 554–559) to account for Milgram-type results. Put a little more specifically, this position requires a specific sort of psychological mechanism—one that tracks the views of other agents and which is connected to the agent's action-producing mechanisms, arguably through emotional mechanisms.

My hypothesis is that Sabini and Silver's social-sensitivity position applies to the trolley and teacup dilemmas. Like the Asch studies, these experiments are about judgments. However, they are about moral judgments, or at least practical ones, instead of judgments of length. Unlike judgments of length, moral codes are inherently for interpersonal regulation. Hence it is reasonable to think that moral judgments involve relatively greater attention to the attitudes and behavior of others than judgments of length. Strong conformity of judgment has been experimentally demonstrated for judgments of length, so it is reasonable to expect to find conformity effects in moral judgments too. Moreover, moral judgments are about values. People care about the things they value, and they are inclined to regulate others' attitudes and behavior toward such things with a wide variety of reactions—e.g., punishment, shaming practices, expressions of resentment. Insofar as it is worth avoiding being the recipient of such responses, it is worth keeping track of both specific and general attitudes about values.

There is an obvious difference between studies of hypothetical moral dilemmas and Asch's experimental protocol. In Asch's study, the subject was in the presence of other people. The subjects in the hypothetical-moral-dilemma studies were alone. One might worry that this undermines

any attempt to extend the study of conformity to these studies from the outset. I have two responses to this worry. First, we should not overstate the solitude of the subjects in the moral-dilemma studies. Although not directly interacting with other people, the subjects were not completely isolated. Other people were present, and, crucially, the subjects were fully aware that they were performing communicative acts by participating in the studies. Subsequent versions of such studies have been performed electronically, so we can assume that many of the respondents were in fact alone. However, they were still performing an act that they knew was communicative and hence other-involving. Second, and more importantly, consider how conformity might be psychologically implemented. Suppose that there is reason to conform one's behavior, regardless of one's judgments, to that of others. One way to do that is to have mechanisms that suppress one's judgments and produce conforming behavior. But another way is to have mechanisms that conform one's judgments to those of others. I see no reason to think that we do not have both sorts of mechanisms. If we had only the first, then the absence of other people would be important for present purposes. But if we have judgment-conforming processes, then the absence of other people is not nearly so significant, as the effects of conformity extend all the way in: our judgments are already subject to conforming pressures. Overall, there is no compelling reason to expect that conformity effects will be completely absent from scenarios such as those found in the moral-dilemma experiments.

There are two conceptually distinct activities at the core of the conformity hypothesis. The first of these is tracking of the views of others, especially their values; the second is effecting of conformity of one's own judgments, views, speech, and/or activity with others in light of the information gathered via the first activity. These activities might be performed by psychologically distinct mechanisms, but they might also be performed by a single mechanism. Emotions are likely candidates for mechanisms that conform agents' views to each other. Sabini and Silver offer embarrassment and confusion as emotional foundations of social conformity. I think fear and shame-aversion can be added. Such emotions might also be suited to tracking of the views of others, which requires some facility with "mind reading" (i.e., understanding the thoughts of others).[14] This is reasonably taken to be a feature of some emotions. We should take seriously the possibility that social conformity is achieved with multiple mecha-

nisms. Some emotions could have distinct mechanisms for mind reading and effecting conformity, whereas others could have single mechanisms that perform both.

In view of the foregoing reflections, here is the conformity hypothesis in the simplest possible form:

(CH) If an action transgresses against prevailing social views, then a person will judge that action to be impermissible.

This is, of course, a *ceteris paribus* hypothesis. Complexities of both individual psychology and prevailing social views will affect whether people judge an action to be impermissible. Of particular importance will be the complexities that come from overlapping or nested groups, and those that come from the fact that individuals can be members of more than one group simultaneously. Nevertheless, these complexities do not affect the experimental cases. The mechanisms by which people are sensitive to prevailing views are posited to be emotional ones, including embarrassment, fear, confusion, shame (and '-aversion' versions to these). Crucially, this is a psychological hypothesis about moral judgment. It is not a normative thesis about what makes something right or wrong. Nor does it imply that people consciously conform to the views of others. People may well be unaware of the degree to which conformity with prevailing views influences their judgments, moral or otherwise. It is reasonable to expect people to neglect conformity in their justifications of their choices. Given that justifications do more than offer accounts of the causes of behavior, such neglect need not be problematic.

On the present view, the explanatory task is simply to explain why the asymmetrical pattern of responses emerges. The reason this pattern emerges might have nothing to do with the answer to the questions why and how we make moral judgments at all. Consequently, the conformity hypothesis abstains from addressing the origins of moral judgment. Instead it addresses the content of such judgments, taking for granted that we make them. This is a marked difference from extant approaches to this topic. The possibility of such an approach is delivered by the general distinction between the psychological origins of moral judgment and subsequent psychological processes of judgment formation and expression. The adequacy of such an approach depends on the empirical details.[15] Although I will emphasize the explanatory power of the conformity hypothesis in what follows, it is

worth keeping in mind the possibility of plural mechanisms here. It could well be that multiple mechanisms do or could produce the asymmetrical pattern of responses.

Let's apply the conformity hypothesis to the trolley and teacup dilemmas. I'll begin with the trolley cases. Here are the cases stripped to the basics:

Let five die or divert the train to save the five but thereby bring about the death of one.

Let five die or push one in front of the train, thereby saving the five but killing the one.

In the context in which these studies have been done—the Western English-speaking world, particularly North America—people are much more critical of direct killing than they are of less direct ways of bringing about death.[16] This is reflected in legal practices of various kinds. Consider as examples the general distinction between murder and manslaughter, the more specific distinction between medical means of directly bringing about the death of patients (all but completely illegal in North America), and the cessation of medical treatment which results in the death of patients (widely accepted in the same medical jurisdictions). If Sabini and Silver are correct about the sensitivity of agents to the views of others, this pattern of views about actions leading to death ought to result in different responses to the trolley cases. The pushing case involves direct killing, which is clearly frowned upon in this social context. But the less direct bringing about of death to save more lives is not clearly represented in this Western worldview.

Significantly, the same kind of explanation applies to the teacup cases. The mother of Billy and Susie offers a directive; this clearly indicates something specific about how she sees the world, including something about what she takes to be valuable. When Billy diverts the train to save the five teacups by breaking one, his immediate action accords with his mother's view as revealed in her directive, in that it is firstly a diverting of the train and only secondly an action that will break a cup. Susie's action, however, is directly a breaking of a cup; hence it reveals a view of the world that is not in accord with her mother's. If Sabini and Silver are correct that we are very sensitive about the accordance of our view of the world with that of others, we should expect the pattern of response to the teacup cases

found by Nichols and Mallon. This marks a difference from the emotion-based explanation offered by Greene and colleagues, which does not apply to the teacup cases.

Nichols and Mallon also present some non-experimental phenomena as problematic for the emotion-based position. Male circumcision, certain acts of war, self-defense, and punishment are personal and emotionally engaging, but are judged to be permissible. The conformity hypothesis fits these cases. Although male circumcision involves full-blooded social interaction, it takes place within a context very much accepting of this kind of action. It is institutionalized culturally, medically, and religiously. The same goes for punishment, with the added feature that at least some of the time the recipient of punishment does not disagree in principle with the meting out of punishment for the offense in question. War too is institutionalized, which signals wide and deep social acceptance. Moreover, the acts of war that are generally judged to be acceptable are ones in which both the aggressor and the victim are combatants, which indicates in many cases bilateral agreement about the general acceptability of such acts. Acts of self-defense involve pragmatic agreement about the acceptability of individual acts of violence between aggressor and defender. That is, the initial attacker signals acceptance of such acts by committing one, and hence the defender acts in conformity with the pragmatically revealed worldview of the attacker by acting violently. This, of course, also typically happens in a context that accepts violent self-defense. Overall, the explanatory power of the conformity hypothesis exceeds that of the original emotion-based explanation (and at least matches that of the rule-based explanation).

The empirical foundation of the emotion-based explanation is the fMRI studies revealing increased neural activity in response to the pushing cases compared to the diverting cases. The conformity hypothesis partially shares this empirical support. Greene et al. found increased activity in Brodmann's areas 9 and 10, among others. These areas have been found to be involved in the processing responsible for embarrassment in response to norm violation, which is exactly the sort of cognitive mechanism posited by Sabini and Silver (Berthoz et al. 2002, 1700). This broadens the empirical credentials of the conformity hypothesis beyond that provided by its links to the body of studies about social conformity.

None of this suggests that people do not have the direct emotional attachment to others that the emotion-based position of Greene and

colleagues posits. Both psychological mechanisms may well be present in normal people. Overall, the present position offers at least a second mechanism to account for the asymmetry in the experimental judgments. Moreover, the conformity hypothesis can offer a partial explanation of the kind of rationale offered by Greene and colleagues. They claim that the personal aspect of the pushing case engages us emotionally in a way that the redirecting case does not. Why is this? One reason might be an emotional mechanism by which we directly care for others. But another reason, suggested by the conformity hypothesis, could be that we are inclined to care about the agreement of our views with those of others. The pushing case calls into question the conformity of our views with prevailing attitudes in a way that the redirecting case does not. Hence it should seem to people that the pushing case is special, but this does not necessarily indicate an inherently altruistic psychological mechanism. Further fMRI tests might shed light on whether both mechanisms are at work—that is, on whether we have functionally distinguishable mechanisms for direct personal engagement and for less direct regulation of our judgments via the prevailing attitudes of a group.

The role of rules in moral cognition is central to the position of Nichols and Mallon. This centrality deserves scrutiny on conceptual grounds. Consider the various phenomena that Nichols, Mallon, and others in the tradition of research on moral/conventional distinction count as rules. In the teacup case, a verbal directive from a recognized authority signals the presence of a rule. If the rule is going to be enforced, it will be enforced by this authority figure. In the trolley case, the rule in question does not stems from such a singular authority but is instead part of the general moral code. It has probably been given a verbal formulation for and by many of the people in the relevant population, but not immediately before the act in question, which is how it happens in the teacup scenarios. Enforcement may be carried out unofficially by any member of this population; for culpable killing, however, there is also an official institution for enforcement. In other studies in this tradition, the sort of phenomenon that counts as a rule has neither a linguistic formulation nor a specific enforcing authority. For instance, in a study of psychopaths and the moral/conventional distinction, Blair groups explicit and implicit reference to rules in justifications by psychopaths of judgments about hypothetical cases as 'normative references' (1995, 15). Smetana notes and accepts this way of

handling these various phenomena in a much-read paper in this tradition: "Social rules may be articulated explicitly by parents and teachers . . . or the law . . . or they may be implicit in children's social interactions (other children may laugh when sex-role expectations are violated)." (1993, 111, examples elided) This is a very diverse collection. Nichols and Mallon posit a specific sort of mechanism, distinct from emotions, that takes as input information about all these phenomena. However, it is worth wondering how plausible it is to posit a special mechanism for both linguistically encoded directives from established authorities and uncodified norms informally enforced by peers. The wider the group, the less likely it is that we have a single psychological mechanism for dealing with it. This is what the rule-based explanation of Nichols and Mallon faces. The more constrained the group of phenomena that count as rules, the more likely it is that we have a single psychological mechanism for handling it. However, such a specific mechanism will not cover the array of cases faced by the rule-based explanation.

The conformity hypothesis has at its core a mechanism that takes a much more restricted domain of input. This hypothesis supposes that the asymmetries in judgment are produced by a mechanism that tracks the values of others. At its broadest, this hypothesis requires mechanisms that can perform mind reading.[17] For the phenomena classified by moral/conventional theorists as rules, I prefer the term 'patterns', or, following Blair, 'norms'. I propose reserving 'rule' for linguistically encoded norms. More importantly, besides the semantic issue there is the substantive issue of the content of the psychological hypotheses. The conformity hypothesis proposes mind reading and emotions as the general psychological capacities that realize our sensitivity to norms. Rules, in the restricted sense, require additional processing for linguistic comprehension. This hypothesis shares with the position of Nichols and Mallon the idea that regularities in social interaction are an important part of the input to moral judgment and thereby figure in generating the asymmetries found in the trolley and teacup cases. However, the conformity hypothesis proposes *no* special mechanism for the array of forms such regularities can take.

It is worth spelling out why norms, including rules, are central to the conformity hypothesis. This hypothesis turns upon the idea that we are psychologically disposed to conform our judgments to the prevailing attitudes around us. When attitudes are prevalent, they form patterns, which

is exactly what norms are. Rules are an especially clear and rigid form in which such patterns can be communicated and enforced. In order to conform to the general views of others, we need to be able to detect, reliably, these patterns of attitudes. That is, we need to be psychologically sensitive to norms.[18]

The conformity hypothesis subsumes the psychological work tapped by the trolley and teacup cases under our general emotional processing. Nichols and Mallon's rule-based explanation posits a separate psychological mechanism for rules. The upshot is that the conformity hypothesis is a simpler one, and insofar as theoretical simplicity is explanatorily desirable, the conformity hypothesis is preferable. More important than this, however, is the relative empirical warrant of these hypotheses. I have already argued for the positive merits of the conformity hypothesis, and we have already seen some of the shortcomings of the rule-based position. What about more direct empirical comparison?

It should be possible to design tests to compare these hypotheses. They might give different results in cases where the subject disavows a rule. Imagine a variation on the teacup cases in which Susie explicitly disavows her mother's rule. In such cases, the rule-based explanation does not apply, so the throwing of the teacup should be as permissible as the diverting of the train. However, under the assumption that the disavowal of the rule does not bring with it complete insensitivity to others' attitudes, the conformity hypothesis should still hold: the throwing of the teacup should seem less permissible because it still reveals a difference in views between Susie and her mother. A more compelling test would be presented in cases in which Susie knows the prevailing attitudes about teacups but there is no rule. This test could be run in one-person and multiple-person varieties. In the single-person case, Susie could know that her mother values teacups highly and hates to see them destroyed, but without her mother's ever issuing any directives about how to treat them. In the multiple-person case, Susie could know that her mother and her mother's friends all love teacups, and hate to see them destroyed, but haven't issued any directives about how to treat them. In these cases, the rule-based explanation should predict no asymmetry between the pushing and the diverting, whereas the conformity hypothesis predicts the same asymmetry that appears in the original versions.

A more compelling test would be provided by modifications to Asch's set-up for studying conformity in judgment. The crucial feature of

Asch's approach is getting people to perform judgments in groups. In the original experiments, the topic of the judgments was length. In the modified version for testing the conformity hypothesis, the topics would be moral, as in the trolley cases, or more broadly practical, as in the teacup cases. Here is a proposed experiment: Subjects would be located in a group of experimental confederates. Both subjects and confederates would be asked to make responses to versions of the trolley and teacup cases; the confederates would go first. For some cases the confederates would give the typical answer. However, for others they would all be instructed, beforehand and without the knowledge of the subject, to give a response that is hardly ever given. The question would be whether the subject gives the usual response or the group response in such a set-up. *Ceteris paribus*, the conformity hypothesis predicts the group response, whereas the rule-based hypothesis predicts the usual response. However, given the wide understanding of 'rule' in the moral/conventional tradition, this variable would be difficult to control. For instance, attention would have to be given to distinguishing emotional mechanisms for conforming to others from rule-discernment and rule-following mechanisms. Moreover, the views of others are made so clearly salient in the Asch-style experimental set-up as to raise worries about environmental validity. In the real world we sometimes encounter the views of others via such direct expressions, but more often we do not.

To finish, suppose that there is a role for rules in normative judgments, and that this plays some role in generating the asymmetries found in the trolley and teacup cases. The conformity hypothesis provides an explanation of why this might be the case. Tracking of norms and rules is important in order to keep track of the attitudes of others; this is important because we interact with others, and their attitudes can have significant effects on our welfare. In contrast, the rule-based explanation offers no account of why rules should be important at all, never mind so important as to generate asymmetries in judgments about cases that involve the same numbers of lives saved and lost.

Conclusion

Without direct empirical testing, it is premature to prefer the conformity hypothesis over the emotion-based explanation of Greene and colleagues or the rule-based explanation of Nichols and Mallon. However, the conformity hypothesis derives support from a variety of sources: its adequacy

as an explanation of conformity effects as studied by Asch and Milgram; its explanatory application to the teacup and trolley cases, as well as to non-experimental cases such as male circumcision, punishment, war, and self-defense; its partial overlapping with the neural support of the emotion-based hypothesis; its ability to shed explanatory light on the phenomena that are central to both of the competing hypotheses.

Moreover, the conformity hypothesis satisfies the three explanatory desiderata offered above. I have also tried to show some of the shortcomings of the alternatives. Given this multi-faceted background, we have as much reason to take the conformity hypothesis seriously as we do the extant emotion-based and rule-based positions. The reasonable position to take is to suspect that multiple mechanisms are at work in generating the asymmetrical pattern.

Even if the conformity hypothesis turns out to be false, it plays an important dialectical role. It shows the possibility of devising explanations of the asymmetrical pattern of responses to hypothetical moral dilemmas that are structurally distinct from the current emotion-based and rule-based positions. This shows that the working assumption—that the asymmetry is to be explained in terms of the psychological origins of moral judgment—is so far unsupported. If something like the conformity hypothesis is correct, then these origins might shed no light on the asymmetry. The silence about secondary psychological processes is not merely an unfilled gap; it is an unrecognized but important challenge to the effect that extant positions have misunderstood their explanatory task.

A particularly important possibility is that the asymmetrical pattern of judgments about trolley cases is produced by entwined judgment and reasoning and/or emotional processes. The assessment of such nuanced possibilities, even though they are offered by the view of moral judgment developed in this chapter, requires new empirical tests. I shall turn to entwined judgment and reasoning in chapter 3 in connection with the phenomenon known as "moral dumbfounding."

Where does all this leave us with regard to our understanding of the asymmetrical pattern of responses to hypothetical moral dilemmas? Pessimistically put, we are no further ahead than we were a decade ago. What looked like progress—the elimination of emotions as playing a central role, in favor of rule-cognition—has turned out to be illusory, owing to a neglect of developments in moral psychology more generally. The important

development in question is the mobilization of the distinction between the origins of moral judgment and subsequent processes of the development and expression of moral judgments. Once we attend to this distinction, the possibility that emotions are responsible for the asymmetry is revived, as seen in the conformity hypothesis. If we take a more optimistic view of the territory, we can claim that it is now marked by the emergence of clearly delineated hypotheses. Moreover, there are lessons here for moral psychology in general: genuine progress in understanding a phenomenon requires not only keeping up with the substantively relevant developments, but also acknowledging that such developments can have implications for the structure of one's explanatory task.

3 Moral Reasoning

3.1 Moral Reasoning: Wide or Narrow?

The topic of this chapter is moral reasoning, by which I mean (following Jonathan Haidt and the tradition of research on moral reasoning stemming from Jean Piaget until the present day, with Lawrence Kohlberg as its prime figure) conscious, intentional transformation of information about moral issues (Haidt 2001, 818). Although it is rarely articulated in this way, the prevailing trend is to assume that our capacities for moral reasoning are located solely within the physical boundaries of individuals. Is this the correct way to think of moral reasoning? The purpose of this chapter is to provide a case for the alternative view: that moral reasoning is widely realized. I shall look at recent work on moral reasoning in the process of making this case. My first job, however, is to convey a sense of how a wide view of moral reasoning might look and some reasons for taking it seriously.

My contention is that moral reasoning centrally and literally (but not solely) takes place between people. Interpersonal contexts provide the resources for the wide cognitive systems in which moral reasoning is centrally realized.

Consider some of the interpersonal jobs that moral reasoning can serve: It provides one with information about the experiences and values of others. Via this information, one's behavior can be attuned to the experiences and values of others—that is, one can deliberately modify courses of action in accordance with what has been learned about others. One's behavior might be regulatable more directly, without deliberate modification by oneself. One's attributions of responsibility can be modified in accordance with information about the experiences and values of others.

Most generally, moral reasoning provides individuals with articulated concepts and patterns of argumentation that have been developed by others, perhaps by themselves or through interpersonal chains of reasoning.

The last job should not be underestimated. Interpersonal moral reasoning consists in the transformation of publicly represented information about moral issues. Language is the most important medium of public representation, but it is not the only one. That such external representational resources as language can greatly augment the cognitive powers of individuals is a familiar theme in externalist treatments of cognition (e.g., Wilson 1994; Clark 2006). For instance, on the basis of considerations of metarepresentation and research programs that examine memory, cognitive development, and folk psychology, Rob Wilson has argued that we should think of higher cognition in general as involving mind-world coupling (2004, chapter 8). "A large part of the significance of mind-world coupling," Wilson writes, "lies in its iterative nature. We take part of the world, and learn how to incorporate and use it as part of our cognitive processing. That, in turn, allows us to integrate other parts of the world that, in turn, both boost our cognitive capacities and allow us to cognitively integrate further parts of the world. And so on." (ibid., 212) Here is how this works for the present line of thought about moral reasoning: Once we become part of the system of moral reasoning (so far only alluded to, not described), we acquire cognitive resources that can be applied not only to other people and to familiar interpersonal topics but also to ourselves and to novel private and public phenomena.

Here is the WMSH account of moral reasoning: Moral reasoning is realized, first and foremost, in an interpersonal moral reasoning system (or multiple systems). The fundamental jobs of this system or systems are interpersonal; they are for bringing about certain sorts of effects on other people. The resources provided by such systems can be used in a self-directed manner by individuals once they have developed the capacities to participate in them, but such use is secondary to its interpersonal jobs. Fundamentally, moral reasoning rests on a platform of social cognition.

Recall the systemicity schema:

_____ systems must be causally and functionally integrated chains of _____ resources, and these, individually and collectively, must play a replicable causal role in _____

The present hypothesis is that moral reasoning systems draw on interpersonal resources that play a replicable causal role in social influence. For any given individual, moral reasoning seems to be primarily a way of interacting with others, and in particular for influencing others. It is only secondarily for bringing about effects on the individual. Influencing one's processes of attributing responsibility or producing actions fall into the class of secondary jobs of moral reasoning, not into its class of primary jobs.

What psychological items must be attributed to individuals for participation in such a moral reasoning system? We must be careful at this point. Certainly the capacities for language comprehension and production are important. So are the various cognitive and affective capacities necessary for general social cognition. However, it is quite possible that the items to be attributed to the individual are not neatly classifiable in such terms. In thinking this I follow Andy Clark's recent hypothesis about public linguistic symbols and human thought (2006). Clark argues that our interaction with such symbols takes a variety of forms, but that one of them in particular deeply extends human thinking abilities. Clark's stalking horse is accounts of language comprehension that require translation of publicly represented information into an inner code. Against this, Clark argues that we have reason to construe public symbol systems as, in some cases, constituting hybrid cognitive systems and ways of thinking with the cognitive abilities we already have (2006, 296–302). Instead of translation into an inner code, very important kinds of cognitive use of linguistically represented information are hypothesized to use the public representations themselves in a cognitively constitutive manner. Clark discusses studies of mathematical cognition (Dehaene 1997; Dehaene et al. 1999) and of simulations of the aid to cognition provided by internal re-use of a public symbol system (Clowes and Morse 2005) as providing empirical reason to take the notion of hybrid cognition seriously. For present purposes, the important point is that the individual psychological items that one needs to participate in such systems need not replicate the resources brought to the table by the public system. What is needed are cognitive capacities for using the public symbols themselves. If this is the case, then the characterization and attribution of such capacities to agents must be done *a posteriori* through careful study of individual-system participation.

These reflections concern the psychological mechanisms required for performance in the moral reasoning system. But it seems conceivable that some people might have such required mechanisms, yet lack deep motivations also required for participation in the moral reasoning system. Why would people participate in such a system? Insofar as such participation comes naturally to us, perhaps this is an idle question. But since there seem to be people outside of this system (i.e., amoralists of various kinds), this is a legitimate question. Perhaps these people lack the performance mechanisms, but perhaps instead they lack something else: the deep motivation that facilitates being included in the moral reasoning system. I do not wish to dwell on this point, but let me point to some potential deep motives, some of them familiar from the long history of philosophical thought about morality: (A) positive feelings about other people; (B) self-regarding positive intrinsic desire for belonging with others; (C) fear of bad effects from not interacting with others in this way; (D) self-regarding fear of not belonging, perhaps to be classified as loneliness-aversion or isolation-aversion; (E) an instrumental desire to succeed in ulterior goals via this sort of interaction with others. Mature humans without something like at least one of (A)–(E) are likely to suffer in reproductive fitness, as they will not be motivated to participate in a deep and important means of interpersonal interaction (assuming that there is no compensating mechanism that correlatively increases reproductive fitness). If this is correct, then we should expect such people to be very much exceptions to the norm. (This is all very speculative, so not much weight should be put upon it.)

So much for the sketch of the WMSH view of moral reasoning. To collect some more fine-grained details relevant to forming and assessing both wide and narrow hypotheses about this topic, let's look at some important research programs that focus on moral reasoning. Special attention will be paid to the emphasis—if any—that these programs place on social interaction.

3.2 Moral Reasoning and Social Interaction

Social interaction has typically been accorded, at best, a secondary role in twentieth-century thought about moral reasoning. It is tempting to see a pattern in the history of moral psychology: Social interaction is emphasized by one theorist, then denied or at least downplayed by the next,

followed by revival and renewed omission. I am inclined to see the overall pattern as vindicating the idea that social interaction deserves a central role in our thought about moral reasoning, but I realize that this may be an interpretation born of personal optimism rather than the history itself.

Jean Piaget and Lawrence Kohlberg

Jean Piaget is one of two figures who cast a long shadow on subsequent studies of moral reasoning. Interestingly, we find both aspects of the suggested pattern in Piaget's work: he conceives of morality and moral reasoning with a focus on the individual, yet cannot avoid letting social interaction into the picture in an important way. Piaget's principal contribution is made in *The Moral Judgment of the Child* (1932; references here are to the 1965 English translation).

Piaget conceived of morality as a system of rules, and hence of moral thought as consisting in understanding and respecting such rules (13). His influential method was to study moral thought by examining its development in children. On the basis of formal studies of children, such as how they learned the rules of such games as marbles, and of observation of his own children, Piaget argued for two stages in the development of moral reasoning. I shall present these in very rough detail.[1] In the first stage, children think heteronomously: rules are received from without and treated as authoritative simply on the basis of their external source. Heteronomous morality is a morality of constraint (197). The natural developmental process is for the child to become suspicious of the trust directed toward these rules and their sources, with the result that heteronomous moral thought is replaced by autonomous thought. In this second stage, which Piaget characterizes as a morality of cooperation (197), the child grows from seeing herself as subservient to the authority of such others as adults to seeing herself as instead as an equal. The challenge here is not to obey the rules one receives from without, but rather to cooperate with others in equitable ways, and hence to find the rules definitive of such fair cooperation (324). Cooperation is, of course, a broad form of social interaction. For Piaget, moral maturity happens through social processes. However, the results of this process are presented as individualistically realized. Through cooperation, rules of reciprocity are interiorized (404). It is not Piaget's contention that mature moral reasoning happens through social processes. It is instead that such processes have effects on the individual

that change her into an autonomous reasoner about morality and the sources of authority of moral rules. The capacities that deliver such autonomous reasoning are narrowly located.

An even bigger shadow is cast by Lawrence Kohlberg.[2] For more than three decades, Kohlberg, in the spirit of Piaget, studied the development of moral understanding. His characteristic method was to ask people about hypothetical moral dilemmas and to examine the sorts of responses that people made when trying to resolve them. Although Kohlberg's work exhibits a recognizable and famous structure across his whole career, I shall focus on the mature formulation found in his 1984 book *The Psychology of Moral Development*. Although Kohlberg continued to speak of the study of moral development, in his later work he circumscribed the scope of his claims to the study more properly of justice reasoning (1984, 224). In the wake of criticism from Carol Gilligan (1982) and others, Kohlberg expanded his notion of morality. Besides justice, characterized by "impartiality, universalizability, and the effort and willingness to come to agreement or consensus with other human beings in general about what is right" (1984, 229), Kohlberg recognized morality as including caring and responsibility with special focus on interpersonal relationships. Since my present interest is in moral reasoning, I shall focus on Kohlberg's work on the psychology of aspects of morality having to do with justice.

Like Piaget, Kohlberg identified stages in the development of justice reasoning. For most of his career, Kohlberg offered a six-stage model of this development. In his 1984 statement of his theory, however, there are two changes worth noting. First, Kohlberg withdrew his endorsement of the sixth stage, thereby officially changing to a five-stage model. Second, he retained the sixth stage and added a seventh, but gave them statuses different from the five official stages of the account. The five official stages are meant to be descriptively adequate (184, 271). They are what Kohlberg calls "hard" stages. Such stages are marked by "discrete operations of reasoning"—i.e., topic-specific, fairly intuitive and non-reflective patterns of thought.

The first two stages of Kohlberg's model constitute what he calls the "preconventional" level (e.g., 1984, 172). Kohlberg claimed that children under 9 years old, and some older people (especially criminals), are at this level (172). Stage 1 of the preconventional level is heteronomous morality, just like that posited by Piaget. Here people treat rules as backed by external

authorities, and they obey these rules out of respect for those authorities and to avoid being punished by them. Stage 2 is characterized by growing individualism and an erosion of respect for external authority. Right is determined by individual interests and perspectives; the agent emphasizes her own perspective while recognizing that others have *their* own perspectives. The next two stages are the "conventional" level of morality (172). Kohlberg claims that this is the level of most adolescents and adults in Western society. Stage 3 is marked by respect for social expectations, especially those of people close to the moral agent. In Stage 4 reasoning, the agent's focal point is on agreements explicitly entered into. After this comes the "postconventional" level, which is instantiated by very few people (172). At this level, social rules are accepted not because of social expectations or agreements but because of an understanding of the moral principles thought to underlie these rules. When social rules conflict with these principles, postconventional agents can reason on the basis of principle rather than social expectation. Stage 5 reasoning is marked by an appreciation of the way considerations of impartiality and the normative significance of idealized rational agents can justify socially specific codes of conduct.

Stage 6 moral reasoning is autonomous in a Kantian sense: moral principles have authority because individuals give these principles to themselves *qua* rational beings. Kohlberg withdrew this from the official model because so few people actually attain this stage. Despite its descriptive inadequacy, Kohlberg retains it as a theoretical postulate that defines an endpoint of the sort of development that his theory concerns (1984, 271). If it were instantiated, it would be a hard stage.

Stage 7 moral reasoning is reflective and hence not a hard stage. Kohlberg calls it a soft stage (1984, 249). Its topics outstrip the narrow focus on justice characteristic of the other stages, but it is not about care and interpersonal relationships either. Stage 7 moral reasoning concerns meaning and the very reason or reasons to be moral. Although he focused most of his research on the earlier hard stages, Kohlberg came to think that this focus was ill-suited to some aspects of adult moral reasoning precisely because of the reflective capacities of adults. Soft stages such as the seventh are needed to account for these distinctively reflective moral concerns (249).

Let's stand back from these details. The Kohlbergian picture of moral development portrays normal moral agency as a process of growth that

slowly yields an increasingly sophisticated individual reasoner. This agent appears to grow rationally self-sufficient; moral maturity, at least with regard to reasoning about justice, is found not in social sensitivity but in cool, deep appreciation of the values that are the foundation for social rules. Kohlberg emphasized this rationalistic individualism in his theory, so it is not unfair to acquiesce in this view of his work. He is famously associated with the multi-stage model of individual moral development. This is a view that seems to be at odds with the externalism of the present book. Yet there are overlooked aspects of Kohlberg's work that soften the opposition between his concerns and mine.

The reader of *The Psychology of Moral Development* may be struck by a section on "sociomoral atmosphere" (1984, 263). To my eye it sticks out like a sore thumb. Its subject is the effects of context on moral thought. Perhaps surprisingly in view of the stages of reasoning that people are thought to go through according to the five-stage model, Kohlberg claims that social context can have significant effects on both the form and the content of moral thought. In one study, Kohlberg found both topic sensitivity and situation sensitivity in moral reasoning. Prisoners exhibited Stage 3 moral reasoning about hypothetical non-prison moral dilemmas but Stage 2 moral reasoning about dilemmas concerning their own prison (264). Such contextual specificity invites wide hypotheses at least as much as narrow ones.

It may also invite pluralistic hypotheses. In a more theoretical vein, Kohlberg (1981, 91) invokes "decalage." The pattern of reasoning characteristic of a certain stage might show up for some topics but not for others. For these others, an earlier stage of reasoning is used. 'Decalage' is the name for the "spread or generalization across the range of basic physical and social actions, concepts, and objects to which the stage potentially applies" (91). Kohlberg discusses decalage in connection with the aims of education, for obvious reasons. It is unsatisfying to teach someone a kind of thought using particular examples only to see the student fail to apply the pattern to other topics to which it is applicable. The prison study just invoked in connection with sociomoral atmosphere is analyzable in terms of a failure of this sort of spread. An attractive hypothesis for this is that moral reasoning is realized by multiple mechanisms that can function in ways that correspond to different Kohlbergian stages. Indeed, Kohlberg holds that the reasoning that is done under the effect of sociomoral atmo-

sphere is distinct from the reasoning that is characteristic of the individual's own moral stage (1984, 265). This is what we should expect if moral reasoning is performed by distinct mechanisms some of which are widely realized.

Although Piaget and Kohlberg cast long shadows, their work has now been overshadowed by subsequent research programs. Because of the importance of this more recent work. I will refrain from wrestling with Piaget and Kohlberg any further.

Reflections on the Moral/Conventional Distinction

Kohlberg laid the groundwork for what has turned out to be the best-developed recent body of empirical studies of moral reasoning: the tradition of study of the moral/conventional distinction, which I briefly introduced in chapter 2. This work has long been central to the Social Domain Theory developed primarily by Judith Smetana (1981, 1993) and Elliot Turiel (1983, 1997). As we have seen, moral/conventional studies are also at the heart of Shaun Nichols's sentimental-rules account of moral judgment (2004a).[3] Note that all these theorists conceive of their projects as concerned with moral judgment. This is not unfair, but these studies are even more clearly about moral reasoning, since they address moral judgment through conscious, intentional deployment and transformation of information about moral issues. Given my defense of the embeddedness of moral judgment in such other psychological processes as moral reasoning in chapter 2, this close connection is exactly the sort of thing that we should expect.

The general method of this approach to the study of moral reasoning is to provide subjects with hypothetical examples of moral and conventional transgressions. Subjects are asked questions about these examples, and their answers are examined for both shared and differentiating notable features. For adults, these examples are typically provided in verbal descriptions. For young children, other means are used. Smetana (1981, 1993) reports studies done with preschool children using pictures of the transgressions. Where necessary, explanations were provided (1981, 1334). R. James Blair reports studies done with school children using Playmobil characters—plastic figures 3–4 inches tall. Standard scripts were used to enact the moral and conventional transgressions (Blair 1997, 189–190). In those studies, positive moral and conventional acts were also enacted, but

most of the literature focuses on transgressions (Turiel 1997, 905). Here are some examples from the Smetana and Blair studies: Smetana asked about such moral transgressions as hitting and not sharing toys. The conventional transgressions in her study included not participating in show and tell and not sitting in a designated place for story time (Smetana 1981, 1334.) Blair focused on potentially harmful moral transgressions such as knocking a child over and damaging another person's bicycle. He asked about conventional transgressions such as wearing inappropriate clothing to school and walking out of a classroom in the middle of a lesson. Positive moral acts included returning a lost toy and donating to charity. Positive conventional acts focused on conforming to social rules about such things as wearing appropriate clothing at school and joining a queue in the prescribed manner (Blair 1997, 189–190). In an overview of the empirical tradition centered around the moral/conventional distinction, Turiel characterizes the moral issues as concerned with physical and psychological harm, and with fairness and justice. In contrast, conventional issues are concerned with social coordination (Turiel 1997, 905; Blair et al. 2005, 57–58).

The results of these studies are striking and interesting. Consistent distinctions between moral and conventional transgressions emerge during the fourth year (Turiel 1997, 905; Blair et al. 2005, 58). Generally, children judge moral transgressions to be wrong even in the absence of rules prohibiting them. In contrast, they treat conventional transgressions as contingent on authority (Blair et al. 2005, 58). Moral transgressions are typically judged to be more serious than conventional ones (ibid., 58). Autistic people distinguish between moral and conventional transgressions (Blair 1996; Nichols 2002, 2004a, p. 10), but psychopaths do not (Blair 1995; Blair et al. 2005, 58–59; Nichols 2002, 2004a). In prison populations, psychopaths treat conventional transgressions like moral ones.

I presented Nichols's sentimental-rules theory of moral judgment in chapter 2. Before Nichols's work, Social Domain Theory was most closely associated with these studies. For instance, Smetana and Turiel are social cognitivists. They think that the core of our moral-psychological abilities consists in our abilities to reason explicitly about moral issues. This is their cognitivism. As a result of their work on children's abilities to discern and reason about moral and conventional transgressions, they argue that there are moral and conventional domains of knowledge that are characterized

by different patterns of reasoning. Children's understanding of these domains is held to be constructed from qualitatively different experiences with kinds of actions and with people with regard to these actions—this is the social aspect of the theory (Smetana 1993, 122; Smetana 2006, 120). This view is supported by observational studies performed with the moral-conventional distinction studies as their background. These studies reveal differing kinds of social interactions, both among children and between children and adults, with regard to moral and conventional issues (Smetana 1993, 122–125). Our early interactions with these different domains are supposed to account for our early abilities to draw the moral/conventional distinction.

The results of the moral/conventional tradition have come under pressure from a variety of studies (Kelly et al. 2007; Haidt et al. 1993; Nichols 2002, 2004a; Nisan 1987; Nucci and Turiel 1993). Instead of reflecting on all of these studies, I will examine only one: Kelly et al. 2007. This is arguably the most important source of pressure on the moral/conventional tradition; it has certainly received the most recent attention. My purposes are twofold: to present the challenge that this study poses to the moral/conventional tradition, and to present a wide hypothesis to explain the resulting picture of moral reasoning.

Kelly et al. start from a curious historical fact: Although the range of studies performed in the moral/conventional tradition has broadened, the content has remained unchanged in one important respect: The earliest studies were performed by developmental psychologists, using young children as subjects. Later studies used adult subjects, including incarcerated criminals and psychopaths. However, the study questions about harm, fairness, and justice retained their juvenile content (Kelly et al. 2007, 121). Kelly et al. took this as their starting point; their study aimed at assessing whether the responses typically found in moral/conventional tradition studies were found when the scenarios presented did not involve "schoolyard" (Kelly et al. 2007, 121) transgressions but instead involved more adult transgressions. Their scenarios included whipping as a punishment on ships (123–124), slavery (124), and corporal punishment in schools (124–125). Their findings are in striking contrast with those of the moral/conventional tradition. Here is just one important point: Subjects judged that prohibitions against harming were not independent of authority (Kelly et al. 2007, 129).

Kelly et al. finish by asking, among other things, why previous research on schoolyard harm transgressions appeared to support the idea that there is a signature response pattern associated with "moral" topics or harm, justice, and fairness that involves judging that rules about these are authority independent, more serious than conventional transgressions, more general, and justifiable by reference to harm, justice, and fairness themselves. They then ask "Is there something special about these simple harm transgressions that is not shared by the more 'grown-up' transgressions that we also used in our study?" (2007, 129). The wide perspective on moral reasoning sketched above gives us some conceptual tools with which to construct a possible answer to these questions.

In particular, my discussion in chapter 1 of clues that suggest that phenomena are apt for the formation of wide hypotheses applies here. First, it is not the simplicity of these transgressions that is special, but their natural home: childhood contexts. The signature pattern of responses associated with the moral/conventional distinction seems to be specific to issues to which children can be expected to be sensitive and with which they will be familiar. Social Domain theorists have long preferred explanations of the moral/conventional distinction that draw on the specific forms of interaction in which children participate. I think that the results of Kelly et al. vindicate this preference, simultaneously circumscribing the scope of this distinction.

Following my discussion in chapter 1, there are two questions to ask about childhood contexts. First, are there cognitive resources specific to these contexts that might be responsible for producing the pattern of responses found in the moral/conventional tradition? The second question turns from context-specific cognitive resources to pressures: Do childhood contexts have unique threats that might give rise to the pattern of responses found in the moral/conventional distinction tradition? Although an affirmative answer cannot be offered without reservation, there is much to think about here. Children are a uniquely vulnerable population among humans. Besides the threats that come from their physical weakness, they are explicitly subjected to lots of instruction, both formal and informal, and to the threat of punishment as part of the correction that goes along with such instruction. Corrective threats take many forms that do not apply to adults: for example, though it is not punitive for my life as an adult to be arranged around periods of time away from other people or

secluded in solitary physical spaces (e.g., my office), children experience these things much less favorably. That is why these can be used as corrective measures with children. Not only are we adults all familiar with such childhood punishments as being sent to a corner or to one's room and being separated from a group of peers by the members of that group; we have been familiar with them since childhood.

Context-specific pressures are more likely candidates than context-specific cognitive resources for explaining the appearance of the moral/conventional psychological pattern in tests in which children are removed from their normal contexts and quizzed about hypothetical cases. It is unclear, to say the least, how contextual resources can play a cognitive role when one is not currently interacting with them. But pressures are different: we have reason, in general, to be careful about environmental threats even when they are not immediately before us. So we can expect the psychological effects of such pressures to be present in tests that remove people from their normal contexts.

What should we make of moral/conventional tests? Do they make any contributions to moral psychology? I think it is reasonable to take the tests at face value, and to grant that they are explicitly designed to assess the structure of moral reasoning. They study the structure of verbal responses and rationalizations given by people who are thinking deliberately and in a self-aware manner about hypothetical examples. This structure includes both the psychological pattern found by the moral/conventional distinction tradition and the context specificity of this pattern found by Kelly et al. It is important to emphasize a difference between the structure and the content of moral thought here. Ordinary people are in touch, "from the inside," with the content of their reasoning about explicitly moral actions. They know, when asked, whether they consider actions of certain kinds to be wrong, and whether they consider actions of one kind to be more serious than actions of another kind. However, they are largely blind to the structural features of their thoughts and their utterances—e.g., the patterns of dependence on authority or transferability that show up in their responses to questions about schoolyard transgressions. These patterns are what the moral/conventional-distinction tests reveal. Since this goes beyond the familiar content of ordinary moral reasoning, it is a genuine contribution. The unique contribution of moral/conventional-distinction testing is the scientific mapping of the contours of our manifest

image of ourselves and others. The problem that arises for Social Domain Theory and for Nichols's sentimental-rules theory is that they place these revelations about our manifest image at the core of theories of moral psychology. But moral psychology is reasonably taken to be about more than our manifest image of ourselves and others: it's also about real-world interactions among people, and very important features of such interactions fall beyond the contours of the manifest image.

Jonathan Haidt

The most important forerunner of the WMSH account of moral reasoning comes from Jonathan Haidt. Haidt draws our attention to two aspects of moral reasoning and overall moral thought that are particularly relevant to present concerns. The first is the interpersonal dynamics of normal moral reasoning. Haidt proposes that moral reasoning be studied first and foremost as an interpersonal process (2001, 814). The natural home for conscious, deliberate transformation of information about moral issues is between people, either through conversation or through other ways of producing and interacting with public representations of the relevant information, such as texts, images, and stories.

Haidt makes this case in part by addressing two sources of bias in reasoning. "Relatedness motives" (2001, 820–821) are sources of bias in reasoning due to relations with other people. Generally, people exhibit a tendency to conform their views to those of the people who are (or are expected to be) nearby. Haidt notes that Darley and Berscheid (1967) found that we tend to judge people to be more likeable when we expect to interact with them than when we have no such interaction, and that Chen, Schechter, and Chaiken (1996) found that people who expected to discuss something with a partner expressed views *before the interaction* "shifted toward those of their anticipated partner" (2001, 821).

"Coherence motives" are the other source of bias. These are caused by mechanisms that serve to defend our views of ourselves and our place in the world from cognitive dissonance. Haidt discusses many kinds of coherence motives. As one example, consider the well-known tendency to seek information that confirms one's own views and to underrate evidence that counts against one's views (Haidt 2001, 821; Haidt cites Baron 1995 and Perkins et al. 1991). If reasoning were produced purely by an aim for truth, we would not fall prey to errors of this sort. However, if there were

psychological pressure to preserve one's pre-existing view of a given topic, this tendency is exactly what we should expect.

What should we make of reasoners characterized by such kinds of bias? Haidt thinks we should take this as indicating that reason is not used solely or even primarily for reflection upon and criticism of one's own views. We reason not like proto-scientists, but like proto-lawyers (Haidt 2001, 820–822): instead of seeking truth, we have pre-existing positions to defend, and we have to deal with others and their pre-existing positions, as in situations of negotiation. For purposes that involve only ourselves, we need not have access to the intellectual resources that are tapped in moral reasoning. Such resources are needed primarily when dealing with others. About his own position, Haidt writes: "The core of the [social intuitionist] model gives moral reasoning a causal role in moral judgment, but only when reasoning runs through other people. It is hypothesized that people rarely override their initial intuitive judgments just by reasoning privately to themselves because reasoning is rarely used to question one's own beliefs and attitudes. . . ." (2001, 819)

Before going further, let's attend to some possible lines of objection. One might think that one's own experience gives lie to Haidt's hypothesis: particular individuals, thinking about their own lives, can know that they engage in moral reasoning mostly privately and hardly at all interpersonally. First, Haidt's hypothesis is consistent with this *occasionally* happening; it is inconsistent only with this being the case for *most* people, and reflection on one's own case cannot pronounce on this wider group. But more importantly, consider the evidence to which one can appeal in making this objection, either from reflection on one's own experience or on the basis of data about a representative sample of a much wider population. The evidence will be reports about the use of a public symbol system, primarily language. As public, the natural home for such a system will also be public. That is, it will be interpersonal rather than intrapersonal. Reliance on evidence that requires such a public system tells as much for Haidt's hypothesis as it does for a line of thought against it. It is difficult to see what else might be adduced as evidence against Haidt; I cannot think of anything. But other sorts of evidence, such as the studies of bias in reasoning, tell in favor of something like Haidt's position. Thus, at the outset, this line of objection, attractive as it may be, does not appear very promising.

A subtler objection focuses on the experimental methods available to psychology. To study moral reasoning, experimenters appear to be confined to studying interpersonal interaction, particularly involving language. Maybe experiments can be designed to gain access to moral reasoning in some other way, but interpersonal exchanges of questions and answers are likely to remain the central tool of inquiry. Given this, we should be wary of assuming that the natural home of moral reasoning is interpersonal. *Of course* this will appear to be the case, but this may be an accidental side effect of psychological methodology rather than an essential feature of moral reasoning.

The answer to this objection brings us to the second aspect of moral reasoning from Haidt that is relevant to the present inquiry. Besides the biases offered by Haidt, so-called moral dumbfounding provides reason to think that moral reasoning *really* is interpersonal. Moral dumbfounding occurs when someone confidently pronounces a moral judgment, then finds that he or she has little or nothing to say in defense of it (Haidt 2001; Murphy et al. 2000). I shall examine this in detail later in the chapter; a full answer to this objection must wait. For present purposes, here is the gist of the idea: When people offer a confident moral judgment and then find themselves without reasons to offer in support of it, this suggests that conscious, deliberate transformation of information about the topic in question is not the immediate cause of the judgment. If such processing ever played a role in coming to this judgment, it has been forgotten. However, for many judgments it is likely that the majority of people have never thought explicitly about the relevant topics. At this point, questions about the nature of subsequent reasoning processes must be answered empirically. Do people tend to come up with reasons for such judgments by themselves? Do they tend to produce reasons via conversational processes with other people? Do they explicitly use the formal resources of public symbol systems? My conjecture is that the answers to these questions show moral reasoning to be *essentially* interpersonal, not just accidentally so. The upshot is that moral dumbfounding indicates the world-involving nature of moral reasoning. I shall defend this position later in this chapter.

Experimental Philosophy

The most recent approach to the study of moral reasoning is found in experimental philosophy. Whereas Social Domain Theory and its central

method (studying the moral/conventional distinction) are well developed and influential, experimental philosophy is nascent. Though its influence is growing, how influential it will be is not yet evident.[4] Where moral/conventional studies have been placed at the foundation of accounts of our overall moral psychology, experimental philosophers have addressed a variety of disconnected topics without yet producing any sort of unifying account of anything. Nevertheless, it is important to examine experimental philosophy. Philosophers in this research program have studied attributions of intentionality, free will and responsibility, moral dilemmas, and the notion of valuing, among other things. Overall, their findings about moral reasoning are interesting.

The starting point for experimental philosophy is philosophers' appealing to intuitions. These appeals often provide a benchmark for philosophical theorizing, which is intended to clarify these intuitions and which, as a consequence, must not distort them. Philosophical appeals to intuition are often cast in terms of what everyone believes, or of what is pretheoretically evident or accepted.[5] That is, the foundation of much philosophical theorizing is ordinary beliefs about certain notions.

Traditionally, philosophers have appealed to their own intuitions as representative of these ordinary beliefs. However, experimental philosophers suspect that this is unreliable, and hence take an empirical approach to determining what ordinary people believe about various notions. Such philosophers design questions so that a particular variable can be controlled and varied, then assess whether there are patterns in ordinary people's responses to these questions.

Joshua Knobe, the pioneer of this philosophical method, has discovered that attributions of intentionality vary in accordance with the moral valence of the consequences produced by the actions in question. People are more likely to say that someone intentionally produced bad effects than that someone intentionally produced good effects, even when the other features of the cases are exactly the same (Knobe 2003, 2006). This is in marked contrast to philosophers' own intuitions about the concept of intention: no theory of intention has ever tied it to the valence of consequences in this manner, or even addressed the possibility of such a connection. More recent work suggests that the same normative sensitivity is found for such other concepts as deciding, desiring, and advocating (Pettit and Knobe 2009).

However, philosophers have guessed correctly about patterns of response to moral dilemmas. The study of trolley cases, discussed in chapter 2,

exemplifies this. Empirical studies have confirmed philosophers' assumption that people will think that pushing someone onto a railroad track to save lives is impermissible, but that switching the track to save lives and inadvertently killing someone is permissible even when the life-and-death math is the same.

Joshua Knobe and Erica Roedder (2009, 132) have recently extended experimental philosophy to the concept of "valuing." As with intentions, evidence suggests that ordinary people are more likely to say that someone's thought amounts to valuing when the content of that thought is morally acceptable or desirable than when it is morally frowned upon.

What do such studies tell us about the place of moral reasoning? Insofar as these studies utilize paper-and-pen tests or on-the-spot answers to questions posed by experimenters, their methodological status is much the same as that of moral/conventional distinction studies. They tell us something about moral reasoning, but we cannot confidently put the results of such tests to foundational use in accounts of the central features of moral agency.

Do these studies reveal anything about their "meta-topic," folk intuitions? Philosophers and psychologists have provided reasons for doubting this. For instance, Peter van Inwagen (1992) and Antti Kauppinen (2007) argue that, contrary to the remarks of philosophers who appeal to them, the intuitions that matter for principled philosophical theorizing are not untutored folk beliefs but other beliefs, such as the beliefs that would characterize the outlook of a rational person considering only the relevant information about a particular topic, or those of informed people who have not made up their minds on the relevant topic. Nahmias et al. (2005, 576) remark that this move makes the according notions into technical concepts rather than folk ones. For some concepts and purposes this is acceptable, but it is not clearly acceptable for such concepts as intention and moral responsibility. Philosophers are legitimately interested in the connections of these concepts to ordinary practices of praising and blaming, and focusing on technical cousins of folk notions does not shed light on these connections.

Another line of reflection is suggested by Jonathan Haidt's work on moral judgment (2001). Recall from chapter 2 that Haidt distinguishes intuition from reasoning. On this view, "reasoning" is the conscious transformation of information, whereas intuition is automatic and unconscious

(ibid., 818). Experimental philosophy addresses moral reasoning, but it does not provide direct information about what Haidt calls "intuition." The important point here is not the semantic one, and nothing substantial turns on whether we think Haidt is correct and the experimental philosophers incorrect, or vice versa, about what "intuition" is. Instead, the important point is that there is a distinction that is important for psychological theorizing, whether it is empirical or more traditionally philosophical. Though philosophers are legitimately interested in what shows up in first-person responses to the questions that experimentalists devise, they can also be legitimately interested in the psychological processes that are not accessible by such introspective methods.[6] These latter sorts of processes may even be part of what philosophers are interested in when they have made traditional appeals to intuition. Insofar as first-person reports cannot gain access to important topics, the move to experimental philosophy addresses only part of what has been of traditional interest to philosophers.

As with studies of the moral/conventional distinction, I think it is appropriate to have the impression that experimental philosophy reveals something about moral reasoning, and perhaps something about our overall moral psychology, but that it is not entirely clear what exactly we are learning. It certainly offers methodological lessons; it is well worth asking what the data and yardstick for philosophical theorizing should be. Its substantial import is less clear. I think we are safe in taking it as analogous to the moral/conventional distinction studies: it charts the structure of moral reasoning, and insofar as this is not evident to us introspectively, experimental philosophy makes a genuine contribution to our understanding of moral reasoning.

3.3 The Theoretical Difference of the Wide Moral Reasoning System

It is time to show how the WMSH view of moral reasoning makes a difference to our understanding of particular topics. The sketch of the nature and cognitive roots of moral reasoning presented above gives a central role to social interaction. I have suggested that the natural home of moral reasoning is interpersonal; that the individual cognitive capacities that enable moral reasoning include those for social cognition and for utilization of interpersonal cognitive resources, centrally including language; and

that moral reasoning is constituted by a cognitive system that extends beyond the physical bounds of an individual into the interpersonal world. Given this, it is reasonable to expect that interpersonal dynamics are at least partly responsible for some of the patterns of moral reasoning found by psychologists and experimental philosophers. In the rest of this chapter, I will develop such a case for moral dumbfounding and for the findings of Joshua Knobe and Erica Roedder on the concept of valuing.

Moral Dumbfounding

Recent interest in moral dumbfounding is due to Haidt and colleagues (Haidt 2001; Murphy et al. 2000). Moral dumbfounding occurs when someone confidently pronounces a moral judgment, then finds that he or she has little or nothing to say in defense of it. Examples of this phenomenon are easy to find. They range from the exotic to the familiar. Some of the imaginary cases typically discussed in connection with moral dumbfounding are rising to near-classic status. Perhaps it is because they are often cases of deviant sexuality. Here is a famous case (Murphy et al. 2000): a brother and a sister, both adults, have consensual sex and use contraception. They do it once, after careful discussion, to see what it is like. The sex is pleasurable. They have a good relationship without sex after this incident, which they remember fondly. Is their incest morally wrong? Many people are confident that it is, but have little to offer in the way of reasons to support such a judgment. However, in my experience as someone who teaches moral philosophy to undergraduates, we need not resort to exotic cases to generate moral dumbfounding, nor should we think that the phenomenon is necessarily linked to strange and unlikely cases. It arises with much more familiar moral issues, whether they are relatively basic or the stuff of newspaper headlines and cable pundit commentary. Is it wrong for one adult to kill another? Is torturing someone for fun morally wrong? Is abortion morally permissible? Cloning? Genetic engineering? My students regularly face such questions in my classes. They regularly have opinions about them, some of them very strongly held. But often they have very little to say in defense of such opinions. Sometimes they have nothing to say at all—it's as if they had never thought about it before, which might well precisely be the case. Perhaps moral philosophy begins where moral dumbfounding occurs. Regardless, the nature and relative prevalence of this phenomenon should be clear.

Moral dumbfounding is one of the phenomena for which Haidt tries to account with his social intuitionist account of moral judgment. More recently, Hauser has also discussed moral dumbfounding in connection with his Rawlsian-creature account of moral judgment. In what follows, I shall first show that Haidt and Hauser assume that moral dumbfounding is to be explained in terms of features of moral judgment that are construed in terms of the intrinsic features of individuals. In contrast, I shall hypothesize that moral dumbfounding is to be explained in terms of moral reasoning, more specifically in terms of social dynamics of such reasoning. Finally, I shall argue that this hypothesis points toward an externalistic account of both moral reasoning and moral judgment. I shall focus on the psychological capacities of moral judgment and moral reasoning or moral reason. Remember, by "moral judgment" I mean the psychological capacity or capacities by which we evaluate actions, states of affairs, and persons in moral terms. By "moral reasoning" or "moral reason" I mean conscious, intentional transformation of information about moral issues. Moral dumbfounding occurs when someone confidently pronounces a moral judgment, then finds that he or she has little or nothing to say in defense of it. The judgment is marked by subjective feelings of confidence that are not rooted in the ability to evince reasons in support of the judgment. We have good reason to take this as a phenomenon associated with moral judgment. It is also one associated with moral reasoning: moral dumbfounding occurs when people try to reason about moral judgments. Despite this connection, both Hauser and Haidt explain dumbfounding in terms of moral judgment. I shall begin with Hauser.

Recall that Hauser draws an analogy between moral judgment and our linguistic capacities, and, on this basis, hypothesizes that we have a moral instinct. (See, e.g., Hauser 2006, 32–42.) Like language, this instinct is constituted by principles to which we do not have introspective, first-person access. This aspect of this model of moral judgment provides a straightforward account of moral dumbfounding (ibid., 156). Just as we effortlessly produce and understand particular languages without being able to offer explanations of the linguistic origins of our utterances, we effortlessly produce and understand moral judgments with analogous ignorance. Often, because of our lack of first-person conscious access, we cannot articulate reasons for the judgments that we confidently make. This is the phenomenon of moral dumbfounding.

Haidt contrasts slow, reasoning processes with rapid, intuitive ones. He thinks reasoning is too slow to account for the rapidity of moral judgment. He explains moral dumbfounding in terms of the intuitive roots of moral judgment (2001, 817). He claims that intuitive processes, besides being rapid and automatic, are not accessible to an agent's subjective, introspective awareness, and that only the results of these processes are accessible (818). This means that subjects can find themselves in touch with the judgments produced by these processes without any access to their source, and hence without anything to say about how they came to have these judgments. Finding something to say about them—i.e., reasoning about them—will take subsequent effort, despite the initial confidence in the judgment that the subject experiences. Overall, this combination of confident moral judgment and lack of things to say in its defense is moral dumbfounding.

There is a pattern to these two explanation of moral dumbfounding. Both ultimately assume that the explanation is to be provided in terms of the processes that generate moral judgments, and that this source is to be found within the physical boundaries of individual agents. I think this pattern deserves scrutiny. I shall examine it by turning to social aspects of moral dumbfounding. Before I do so, let me note that Prinz gives an account of moral dumbfounding that takes greater notice of its combination of moral judgment and moral reasoning than the accounts of Haidt and Hauser. Prinz's account of moral judgment has room for reasoning at the first stage (that of categorizing an action or a state of affairs). Working backward from an expressed judgment, the sorts of reasoning that are involved at this early stage might be reported as justifications of how one feels. Alternatively, and more in the spirit of Haidt, they might be offered as *post hoc* rationalizations even if they were not included in the process that produced the judgment. About some feelings, however, we will not be able to say anything: they will be "basic values" for us (Prinz 2007, 32). For Prinz, cases of moral dumbfounding involve basic values. Though Prinz's account of moral dumbfounding explains it in terms of both moral judgment and moral reasoning, it still exhibits the individualism found in the accounts of Haidt and Hauser. Let us now turn to social aspects of moral judgment and moral reasoning, to see whether a persuasive and wide account of moral dumbfounding can be produced.

Social Aspects of Moral Dumbfounding

Moral dumbfounding is strikingly similar to a phenomenon that has been noted by Robert Wilson and Frank Keil (Wilson and Keil 1998; Wilson 2004) in connection with causal explanations. People are often confident that they know how something (e.g., a car, a toilet, a computer) works, but when asked for explanations they often find that they have little or nothing to say about how the device in question works. Wilson and Keil call this the "shallows" of causal explanation (Wilson and Keil 1998, 147–158; Wilson 2004, 202). When conjoined with the ordinariness and "ubiquity" of explanations, the shallowness of explanation generates an apparent paradox: how can something that is beyond the reach of ordinary people be so common? Wilson and Keil argue that a division of cognitive labor dissolves the paradox and explains the appearances. We rely for explanations on the knowledge of others, and our assumption that others have this knowledge gives rise to our confidence that we know how things work (Wilson 2004, 204). What we really know is that we can easily get information about how something works; the confidence that rests on this foundation is spread onto our sense of our own grasp of the causal workings of the world at large. As a consequence of the availability of the knowledge of others for explaining things, we are not required to carry around a lot of internally stored information about the world. Day-to-day facility with using things and the know-how to access others' knowledge about the causal depths of the world are all we need to do very well in handling the world, both when it functions normally and when we need information about its normally hidden depths in order to repair it (Wilson 2004, 204–205).

This line of thought applies to moral reasoning. One of the things we do with moral reasoning is offer explanations. Indeed, there is a large literature in meta-ethics about the pros and cons of moral explanations relative to scientific ones.[7] Insofar as the intellectual endeavor of offering moral explanations is analogous to that of offering causal explanations, the illusion of explanatory depth discussed by Wilson and Keil is to be expected in the moral domain. This seems to be moral dumbfounding. With regard to moral issues, all we need to get around successfully in the world are superficial (shallow) knowledge of what is wrong, what is right, and what is permissible and knowledge of how to get more information about these things when we need it.[8]

In the causal case, deep explanatory knowledge is provided by experts. If I want to know about chemical reactions, I can ask my father, a chemistry professor. If I want to know about something in my garden, I can ask my mother, a botanist. If I want to know about repairing my toilet, I can ask my father-in-law. My father-in-law isn't a plumber; he is an electrician who happens to have lots of practical know-how about household things. I didn't appreciate this until I bought a home; now I appreciate it a great deal. If none of these people is available, libraries and the Internet provide lots of readily available information about the causal innards of such things as toilets and linden trees.

Something similar holds for moral issues. People can, in principle, identify moral experts of various kinds. Professors, pundits, and clergy can all lay claim to being moral experts, insofar as such a job description can be filled. These people are analogous to the chemists and botanists of the causal case. Parents and worldly people in general are the moral analogues of the non-formal expert exemplified by my father-in-law. Books and other repositories of information containing lots of information about moral issues are readily available to ordinary people. So long as we can rely on such sources of information, we generally do not have to carry the details of moral issues or theories around with us. We can turn to other people when the need arises. Thus, we should expect to find ourselves with little to say about what vouchsafes our confidence in particular moral evaluations from time to time. This does not require that these people are seen as infallible sources of moral authority. This line of thought is not confined to explanations of the for, "X is wrong because A says so." It does not even give a central role to such appeals to authority. Instead, the moral experts here are, first and foremost, *exactly* akin to experts on botany or plumbing—people who have knowledge about the details of phenomena within a particular domain. A plumbing expert can provide plumbing explanations while playing no role in their content. The same goes for moral experts.

The present account of moral dumbfounding differs from the explanations offered by Hauser and Haidt in the following ways:

Haidt and Hauser explain moral dumbfounding in terms of the processes that generate peculiarly moral judgments. The present account instead locates moral dumbfounding in a broader context of reasoning patterns characteristic of the provision of explanations generally.

Hauser and Haidt (and Prinz) explain moral dumbfounding in terms of individualistically located processes that are introspectively inaccessible. The present account instead offers an externalistic account in terms of the reliance on the knowledge of other people as cognitive resources.

Prospects for Empirical Assessment of the Social Dependence Hypothesis

At this point, progress depends on empirical assessment. Without data we cannot decide between the individualistic and wide hypotheses. We cannot even know whether there is a single source of moral dumbfounding or whether we should be pluralists about it. Here are some suggestions about the sort of study that may be useful. These empirical speculations will also serve to refine the present hypothesis.

The natural things to assess are points of contrast. First, whereas Hauser and Haidt explain moral dumbfounding individualistically, the present hypothesis explains it in terms of our reliance on others' knowledge as a cognitive resource. Second, Hauser and Haidt locate the roots of moral dumbfounding in the sources of moral judgment. In contrast, the present hypothesis explains it in terms of moral reasoning. I shall begin with moral reasoning, and then turn to social conformity.

Moral Reasoning

The social-dependence hypothesis about moral dumbfounding has at its core the apparent similarities between offering causal explanations and offering moral justifications. One reason for this is the apparent similarity between moral dumbfounding and the so-called shallows of causal explanation. It is worth looking more closely at causal explanations and moral reasoning to see whether there are other similarities, whether there are dissimilarities, and, most importantly for present purposes, whether studies of the shallows of causal explanation can pave the way for studies of moral dumbfounding.

Since formulating the shallows-of-causal-explanation hypothesis with Rob Wilson, Frank Keil and colleagues have assessed it empirically. Their studies confirm that the so-called illusion of explanatory depth occurs. They also shed light on factors that contribute to that illusion, and provide reason to be wary of generalizing from this phenomenon to others. Since I am making an extension of this kind, I shall begin with this point.

Rozenblit and Keil (2002) contrasted causal explanations with knowledge about facts, procedures, and narratives.[9] They found that subjects

were most prone to misunderstanding their own knowledge about explanations; they were much less likely to misconstrue their knowledge of facts, procedures, and narratives (ibid., 18–27). On this basis, they caution against studying "overconfidence" as a general phenomenon that is insensitive to differences between domains. For present purposes, the important question is whether the kind of moral reasoning that is used in constructing and offering moral explanations is most similar to causal reasoning or to one of the other domains studied by Rozenblit and Keil. In general, there is good reason to see the construction and offering of moral justifications as most similar to causal explanation. When offering a moral justification, you are typically not (merely) showing that you know a fact. For example, explaining why you think it is wrong to kill one person in order to use that person's organs to save the lives of five others is not particularly similar to showing knowledge of provincial capitals. Nor is it much like demonstrating that you know how to do something—for example, to run a formal meeting in an office. Again, it is not much like showing that you can follow a narrative (who did what to whom, when, where, and so on). Instead, in offering a moral justification you are following and displaying the logic of moral values. You show how values figure in a particular situation or kind of situation. This involves showing which values are relevant and, of these, which is more important than another, and why. The mechanical metaphor of laying out the moral "parts" of the situation, and of showing how they work together, strikes me as apt. Such a metaphor connects this sort of moral reasoning to causal explanations, not to knowledge of facts, procedures, or narratives. Thus, *prima facie*, the extension of the shallows of explanation hypothesis moral reasoning is reasonable.

Rozenblit and Keil suggest some features of causal explanations that they think are responsible for eliciting overconfidence in one's knowledge of them. These features both give substance to the comparison of moral reasoning with causal explanation and suggest ways of assessing the social dependence hypothesis regarding moral dumbfounding. Here are four of the features they offered (Rozenblit and Keil 2002, 2–3, 34–38):

(A) When people successfully interact with devices with perceptually vivid parts, they misunderstand how much their success is due to their abilities to interact with the environment as opposed to retrieval of internally represented knowledge.

(B) Explanations of causal phenomena can be complexly hierarchical, and people misconstrue familiarity with one level for understanding of workings at other levels.

(C) Explanations have indeterminate endpoints, such that knowing how something works up to a point can be misunderstood for complete knowledge of the relevant workings.

(D) Explanations are rare.

Suppose that Rozenblit, Keil, and others are correct that these features of causal explanations are at the heart of the shallows-of-causal-explanation phenomenon. Do they pertain to moral reasoning? *Prima facie*, (B)–(D) all apply to moral explanations. They are relatively rare. It is easy for moral philosophers to overestimate how common they are. However, teaching moral philosophy to undergraduates gives me the impression that such reasoning and, especially, interpersonal discussion are fairly rare. My students are only sometimes in the position of being required to say why they think something is right or wrong. The relevant skills typically must be developed in university, as they do not get much attention earlier in an individual's education.

(B) and (C) are also found throughout the domain of moral justifications. It is one thing to know that (to continue with the example) killing one patient for organs to save five other patients is wrong. Part of knowing why this is wrong has to do, in all likelihood, with respecting human life, which many people understand. But, e.g., explaining what makes respecting humans important, and why it is one value rather than something else that is important in this case, and why these considerations do not justify killing one patient to save five others, is quite something else. Understanding the general case about human life does not entail understanding the logic of the relevant values. This example also exemplifies the indeterminacy of the endings of moral explanations. In some senses citing the value of human life explains the initial judgment adequately; in other senses it barely scratches the surface.

(A) is different. It is also very important, since Rozenblit and Keil think it is the most important factor in predicting the occurrence of the illusion of explanatory depth. At first glance it seems not to apply to moral judgments and justifications. These do not have perceptually vivid parts. However, at second glance there are more similarities here than there

initially appear to be. Moreover, this feature of Rozenblit and Keil's discussion of causal explanations points to ways of studying moral dumbfounding empirically.

Our interactions with cases that elicit moral judgment and moral reasoning take a variety of forms. Here are just two differentiating distinctions: First, such thought is provoked in cases that we actually experience. (Example: You are in a hurry, but you come across somebody who needs medical assistance. What should you do?) But other cases, especially those used in recent studies of moral judgment, are merely imagined. Second, some cases calling for moral thought are very concrete, whereas others are abstract.[10] In principle, these distinctions are orthogonal to each other. (For an illustration of four categories of cases, see figure 3.1.)

It is clear that imagined cases can be either abstract or concrete. One might think that actually experienced cases must be concrete, but there is reason to think that we can experience different degrees of abstraction. There might be a difference between seeing somebody in front of you who needs medical attentions (a very concrete case) and experiencing a case of injustice due to systematic inequalities in voting laws. The latter strikes me as requiring significantly more abstract thought about moral values.

One important question to ask is whether all four kinds of cases elicit moral dumbfounding; a second question is, if so, whether they all do so equally. If constructing and offering moral justifications really is very

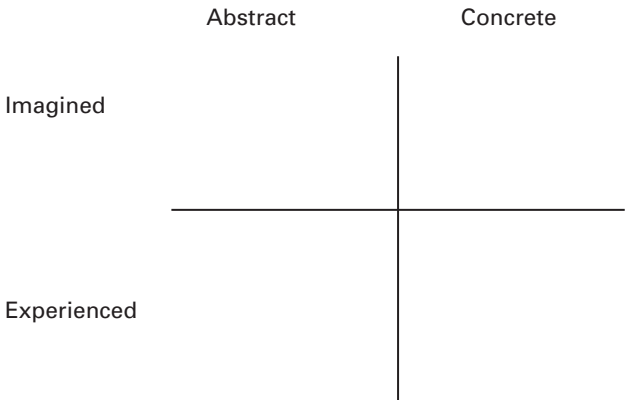


Figure 3.1

similar to devising causal explanations, there is reason to expect the concrete cases to elicit moral dumbfounding significantly more than the abstract ones, and the actually experienced concrete cases to elicit it most of all. This is because of (A) in Rozenblit and Keil's list—the point that dealing with devices with perceptually vivid parts is the most important elicitor of illusions of explanatory depth. "Perceptual vividness," or its moral analogue, is greater in concrete cases than in abstract ones, and greatest of all in actually experienced concrete cases. These predictions should be at least partially open to empirical investigation; certainly varying imagined cases in terms of concreteness and abstraction is possible. Studying actual moral responses to scenarios is more difficult, but not impossible. Studying people's reports of past experiences is one way. Another is to consult, or to perform, studies of the sort characteristic of the person-situation debate.[11] Two interesting and more specific questions generated by this taxonomy are which distinction, if either, is more significant in eliciting moral dumbfounding, and whether experienced-abstract cases are more likely to elicit moral dumbfounding than imagined-concrete ones or vice versa.

Actually experienced concrete cases raise issues that open still more avenues to empirical research. Perhaps more than the others, this sort of case is likely to be connected to other aspects of a person's life. The justifications offered in such cases are accordingly more likely to cite particular facts about a person's life. Consider this example: "Why did you think that X was the right thing to do?" "Well, Bob is my friend, and a few weeks ago he helped me out with some problems." That is, some justifications in the face of actually experienced concrete cases are going to take the form of narratives. Keil et al. (2004, 236) argue that the abstract nature of causal explanations, as opposed to the particular details of narratives, is one of the reasons we overestimate our causal explanatory knowledge more than our understanding of narratives. If I am correct about actually experienced concrete cases, some empirical predictions can be made. First, I predict that narrative justifications should be found significantly more often for experienced concrete cases than for the other types. Second, on the basis of the findings of Keil and colleagues, I predict that moral dumbfounding will be elicited less often in cases in which agents can easily offer narrative explanations. This is an interesting complication in view of the previous observation that. since actually experienced concrete cases are most like the

perceptually vivid experiences that give rise to illusions of causal explanatory depth, we are most likely to find moral dumbfounding here. The combination of these lines of thought yields a subtler prediction: that moral dumbfounding is, overall, most likely to be elicited by actually experienced concrete cases for which no narrative justification is easy to offer, and much less likely to be elicited by actually experienced concrete cases that connect with other aspects of an agent's life and for which narrative justifications can easily be offered.

Since these are empirical issues, not much in the way of concrete conclusions can be drawn before data are collected. However, it is reasonable, on the basis both of these reflections about moral thought and of the findings of Keil and colleagues, to expect to find *multiple* sources of both moral judgment and moral reasoning, and hence of moral dumbfounding. I shall have more to say about such pluralism in chapter 6. The present point is that the more some of these sources of moral judgment and moral reasoning are like our thought about causal explanations, the more support there is for the present hypothesis about moral dumbfounding.

Social Dynamics

One might assess the social-dependence hypothesis regarding moral dumbfounding indirectly by studying cultural variation in moral thinking. Such studies are indirect because they do not study moral dumbfounding itself. However, this sort of information is relevant because the extant individualistic and social accounts of moral dumbfounding seem to make different predictions about cultural variation in moral thinking by virtue of their more general features. The present hypothesis accounts for moral dumbfounding in terms of under-recognized social dimensions of moral reasoning, and it detaches moral dumbfounding from the individualistic sources of moral judgment emphasized by Haidt and Hauser. *Prima facie*, it derives general support from evidence of cultural variation in moral thought, whereas the extant individualistic accounts derive general support from evidence of patterns of moral judgment that are independent of moral reasoning.

There is evidence for both general patterns. Hauser et al. (2007) have found evidence of patterns of moral judgment that are not accounted for by the information represented in the justifications offered for these judgments. This points toward sources of moral judgment independent of

moral reasoning. However, a variety of studies have found evidence of cultural variation in moral thought.[12] Proponents of the "CAD" account of moral thinking group values in three categories:

Community This group "relies on regulative concepts such as duty, hierarchy, interdependency, and souls. . . . It aims to protect the moral integrity of the various stations or roles that constitute a 'society' or 'community', where a 'society' or 'community' is conceived of as a corporate entity with an identity, standing, history, and reputation of its own." (Shweder et al. 1997, 138)

Autonomy This domain "relies on regulative concepts such as harm, rights, and justice . . . and aims to protect the discretionary choice of 'individuals' and to promote the exercise of individual will in the pursuit of personal preferences." (Shweder et al. 1997, 138)

Divinity, which "relies on regulative concepts such as sacred order, natural order, tradition, sanctity, sin, and pollution" (Shweder et al. 1997, 138).

Turiel et al. (1987) criticized early work in favor of the CAD taxonomy of the moral domain, but subsequent studies taking into account these criticisms—especially Haidt et al. 1993; see also Rozin et al. 1999—have provided support for it. Crucially, there seems to be cultural variation in emphasis among the three domains, if not in initial predispositions that enable children to classify values in these ways or in the mere cultural presence of thought about all three categories.

More direct assessment of social aspects of moral thought might be accomplished through adaptation of Solomon Asch's approach to studying conformity in other sorts of judgment. I discussed this sort of study in chapter 2. The crucial methodological feature is that this experimental protocol places a subject in a group of experimental confederates. The members of the group are asked to make judgments of various kinds. For some questions the group answers predictably, but in other cases the confederates give surprising answers. This protocol can be used to assess the extent to which the subjects' reported judgments conform to the group pattern (Asch 1951, 1952, 1955, 1956; Ross and Nisbett 1991, 30–32).

There are several axes along which Asch's protocol could be modified to assess the social aspects of moral thought in general and moral dumbfounding more particularly. First, Asch's own set-up could be used, but with

moral judgments in place of judgments of length. Discovering social conformity would provide more support for the present hypothesis than for rival individualistic ones. Second, this set-up could be varied to include purported moral experts. The simplest way of doing this would be to have one subject in a group of experimental confederates, with one playing the role of moral expert. After some easy cases, the expert would confidently make (and perhaps defend) a very odd moral judgment. Whether the other confederates concur or not could be systematically varied. Whether the subject concurs with the expert, or with nobody, or with confederates who disagree with the expert would provide information about the sorts of social dynamics at work in moral thought (if any).

Third, Asch's protocol could be modified to apply directly to moral dumbfounding in particular. This could be done with either expert or no-expert configurations, and it could take two different forms. In one form, the group could be presented with a case known to be likely to elicit moral dumbfounding (perhaps on the basis of studies like those sketched in the preceding subsection), and they could attempt to resolve it through discussion. In the expert set-up, one confederate would present and argue for explicit resolutions. After some fairly straightforward cases, the expert would defend an odd justification. Whether the group concurred or not could be systematically varied. Whether the subject concurs with the expert, or with nobody, or with confederates who disagree with the expert would provide information about the sorts of social dynamics relevant specifically to moral dumbfounding. In the no-expert set-up, the group as a whole would attempt to resolve the cases, with different confederates leading the way for different cases. After some easy cases, the group could offer an odd resolution to a case. Whether the subject concurs would provide information directly about possible social aspects of moral dumbfounding.

In a second form, instead of presenting subjects with cases known to elicit moral dumbfounding, the point would be to assess whether moral dumbfounding could be made to happen. Again, this could be run in both expert and no-expert configurations. The group would be presented with straightforward moral scenarios, and the confederates would give the same answers. After a few cases, either the whole group or a leading expert could feign moral dumbfounding. In the expert scenario, whether the other confederates concurred could be varied systematically. Again, whether,

with whom, and to what degree the subject concurred would illuminate the social dynamics (if any) of moral dumbfounding directly.

Implications for the Study of Moral Reasoning and Moral Judgment

I shall conclude this section by briefly gesturing toward implications for thinking about moral reasoning and moral judgment. Recall that Haidt describes moral reasoning as an essentially interpersonal process. It is worth emphasizing that the present account adds an independent line of reasoning to this case. Haidt makes this claim primarily because of studies of moral thought, such as responses to questions about moral and conventional transgressions. (See, e.g., Haidt et al. 1993.) The present argument works by locating moral reasoning, and moral dumbfounding, in a broader context of thought about explanatory reasoning and observation of the shallows of explanation. As a multi-faceted case for this possibility develops, this approach to moral reasoning should be taken more seriously in moral psychology.

Trickier implications arise with regard to moral judgment. If the present hypothesis about moral dumbfounding turns out to be correct, then complications arise for the explanatory pattern exhibited by such positions as those of Haidt and Hauser. Most simply put, we could not directly infer from thought about moral dumbfounding that the sources of moral judgment account for it, or that moral judgment should be construed individualistically. To construe moral judgment individualistically would imply what I shall call the significant *psychological independence* of moral reasoning from moral judgment, given that the present considerations point toward an externalistic model for moral reasoning. And given that, as described, moral dumbfounding is a phenomenon that shows up where moral judgment and moral reasoning meet, such independence should be directly defended. Perhaps such argument can be provided, but it has not been presented yet. Assessment of the likelihood of this case cannot be performed here. Instead, let me point to two other theoretical possibilities.

Following chapter 2, one possibility is the embeddedness of moral judgment in moral reasoning. I mean this in a closer sense than that implied by Haidt's social intuitionism. Instead of the sources of moral judgment being located subconsciously within the physical boundaries of individual agents, perhaps they are instead widely distributed among people. In this

case, the very same social features that constitute moral reasoning and account for moral dumbfounding would also serve as an important part of the source of moral judgment.

Again following chapter 2, another possibility is pluralism about moral judgment. This yields the possibility that we have, simultaneously, psychological sources of moral judgment that should be individualistically construed and that are independent of moral reasoning, and other sources of moral judgment that should be widely construed and that are entwined with moral reasoning. This, of course, raises the possibility of plural sources of moral dumbfounding. Perhaps some forms are rooted in social aspects of moral reasoning whereas others are brought about by individualistic processes of moral judgment. In a sense, perhaps both the individualistic hypotheses of Haidt and Hauser and the present social-dependence account can be true of normal humans.

Issues of pluralism and non-pluralism, of individualism and externalism, and of independence and embeddedness must be decided empirically, so I cannot offer a decisive case for one view over the others. However, here are two sorts of consideration that lend support to views that treat moral reasoning and moral judgment as entwined.

First, consider the long-standing research programs that are construed, by their practitioners and by others, as assessing the nature of moral judgment. Arguably the most significant of these is the research done on moral and conventional rules and transgressions. Consider the way these studies work: Generally, subjects are provided with hypothetical examples of moral and conventional transgressions.[13] They are asked questions about these examples, and their answers are examined for both shared and differentiating notable features. It should be clear that in such studies not only are subjects asked about the categorization of events in moral terms (i.e., asked to perform moral judgments); they are simultaneously asked to perform moral reasoning. There is, so far as I can tell, no research program that studies moral judgment without moral reasoning, and this raises the question of whether doing so is even possible. The practical difficulties of disentwining moral reasoning and moral judgment lend *prima facie* support to theories that preserve these deep connections.

Second, recall some of the interpersonal jobs that moral reasoning seems to serve:

It provides one with information about the experiences and values of others.

Via this information, one's behavior can be attuned to the experiences and values of others. That is, one can deliberately modify courses of action in accordance with what has been learned about others.

More generally, moral reasoning provides individuals with articulated concepts and patterns of argumentation that have been developed by others, perhaps by themselves or through interpersonal chains of reasoning.

I think it is reasonable to see moral judgment as entwined in these interpersonal jobs of moral reasoning. Arguably, what one gains access to via interpersonal moral reasoning is a powerful means of categorizing actions in moral terms. This suggests another source of moral dumbfounding: We can be confident in our moral judgments without having immediate access to chains of moral reasoning not only because of a cognitive division of labor in which we rely on other people, but also because we rely on access to an interpersonal system of moral reasoning that can be used as a tool to generate answers to moral questions. Moral reasoning requires the transformation of information about moral issues, which can be time-consuming and difficult. Moreover, if Haidt is correct that the natural home of moral reasoning is interpersonal, then individuals asked to perform moral reasoning alone, in response to questions directed solely at them, might well find themselves using this powerful tool in an unfamiliar and second-best fashion. All of this is suggestive at best, but what it suggests is a picture of moral judgment and moral reasoning as deeply involved in each other and extending into the world beyond the physical boundaries of individual agents. This way of thinking about moral judgment and moral reasoning has been unduly neglected in recent work on these topics in general and on moral dumbfounding in particular.

3.4 Joshua Knobe and Erica Roedder on the Folk Concept of Valuing

In my brief discussion of experimental philosophy, I mentioned work by Knobe and Roedder on the concept of valuing. 'Valuing' is the name of a kind of psychological stance that people take toward objects, ideas, states of affairs, and so on. Generally, one can like something without valuing

it. Knobe and Roedder claim that evidence suggests that application of the folk notion of valuing to a person is sensitive to the moral value of the object or state of affairs that is the content of the person's thought. In the rest of this chapter, I shall review Knobe and Roedder's data and suggest that a variant of the social-sensitivity hypothesis provides just as good an explanation.

To determine whether the moral valence of the object of one's thought affected judgments about whether someone valued that object, Knobe and Roedder devised studies that presented people with hypothetical cases. In their first study, subjects were presented with a pair of cases. These cases were designed to be identical except for the value of the object of thought. Here are the studies in their original detail (Knobe and Roedder 2009, 133–134):

Case 1: George lives in a culture in which most people are extremely racist. He thinks that the basic viewpoint of people in this culture is more or less correct. That is, he believes that he ought to be advancing the interests of people of his own race at the expense of people of other races.

Nonetheless, George sometimes feels a certain pull in the opposite direction. He often finds himself feeling guilty when he harms people of other races. And sometimes he ends up acting on these feelings and doing things that end up fostering racial equality.

George wishes he could change this aspect of himself. He wishes that he could stop feeling the pull of racial equality and just act to advance the interests of his own race.

Case 2: George lives in a culture in which most people believe in racial equality. He thinks that the basic viewpoint of people in this culture is more or less correct. That is, he believes that he ought to be advancing the interests of all people equally, regardless of their race.

Nonetheless, George sometimes feels a certain pull in the opposite direction. He often finds himself feeling guilty when he helps people of other races at the expense of his own. And sometimes he ends up acting on these feelings and doing things that end up fostering racial discrimination.

George wishes he could change this aspect of himself. He wishes that he could stop feeling the pull of racial discrimination and just act to advance the interests of all people equally, regardless of their race.

After reading case 1, subjects were asked whether or not they agreed with the sentence "Despite his conscious beliefs, George actually values racial

equality." After reading case 2, subjects were asked whether or not they agreed with the sentence "Despite his conscious beliefs, George actually values racial discrimination." As it turns out, subjects were significantly more likely to say that George valued racial equality than that he valued racial discrimination.[14]

In a second study, Knobe and Roedder used a single case but presented it to two groups with distinct ideas about the values in question. Specifically, they ran a case about premarital sex by one group of random passersby in New York's Washington Square Park and by another group composed of participants in a Mormon Bible-study group. Most of the members of the first group said that refraining from premarital sex was neutral; most of the second group said it was good. Here is the case with which they were presented:

Case 3: Susan grew up in a religious family, but while she was in college, she started questioning her religious beliefs and eventually became an atheist.

She will be getting married in a few months to her longtime boyfriend. Recently, the subject of premarital sex has come up.

Susan definitely has a desire to have sex with her boyfriend, but whenever she thinks about doing so, she remembers what her church used to say about premarital sex and feels terribly guilty. As a result of these feelings, Susan has not had sex yet.

Because she is no longer religious, Susan believes there is nothing wrong with premarital sex. She wishes she could stop feeling guilty and just follow her desires. (Knobe and Roedder 2009, 135)

Subjects were asked whether Susan valued refraining from premarital sex. Most of the park-goers said that she didn't; most of the Bible-study participants said that she did.

Knobe and Roedder claim that their studies provide evidence that the moral value of the objects of people's thoughts affects folk ascriptions of the concept of valuing. However, another explanation is warranted by the data that have been collected. That explanation is that judgments about whether someone values something are sensitive to prevailing views about the value of the object, idea, or state of affairs in question. Here is the second social-sensitivity hypothesis, in a simple form:

Ascriptions of the folk concept of "valuing" are sensitive to prevailing social views in a particular context. Thus, if an object, a state of affairs, or an idea is not valued in the prevailing social views in a particular context,

people in that context will tend to judge that people do not value it. If an object, a state of affairs, or an idea is valued in the prevailing social views in a particular context, people in that context will tend to judge that people value it.[15]

There are at least two reasons why it is important for people to track the values of others, and hence why such sensitivity would show itself in studies of moral reasoning about values. First, there is the important and simple matter of reasoning about generalizations: if asked what a given person values, the prevailing views about values in the given context are the relevant background on the basis of which to predict the values of that given person. Second, and more closely related to present concerns, if one is interested in influencing others and in regulating one's own behavior to fit in with others, having information about their values is very useful.

The George cases were presented to people in a Manhattan park. In this context, the prevailing views are much more likely to value racial equality than discrimination, so the social-sensitivity hypothesis finds support here.

For comparing the social-sensitivity hypothesis with that of Knobe and Roedder, the second case is more telling. There is reason to think that it provides more direct support to the social-sensitivity hypothesis than to the hypothesis offered by Knobe and Roedder. The reason is that Knobe and Roedder fail to distinguish clearly between the actual value of ideas, objects, and states of affairs and prevailing views in particular contexts about these things. They formulate their hypothesis in terms of the actual value of the objects, etc., but their tests, especially the second one, turn upon differences in prevailing views about what is valuable. Since the second test is designed explicitly around differences in views between two groups, it provides fairly direct support for the social-sensitivity hypothesis and less direct support for Knobe and Roedder's hypothesis.

Can tests be designed to compare these hypotheses directly? Doing so might seem to require a case in which the actual value of something is demonstrably different from the prevailing views about that object in a given context. This is a very difficult arrangement to ensure. However, the Susan case provides *prima facie* reason to think that we do not need to design such a test. If application of the concept of valuing is to be sensitive to the actual value of objects, ideas, and states of affairs, and not to prevail-

ing views about the values in question, then agents must have mechanisms, accessible to moral reasoning, for tracking such values that operate independent from those that track prevailing social views. The Susan case provides reason to think that we do not have such mechanisms. In this case, the crucial issue is the value of refraining from premarital sex. Presumably it cannot be both neutral and not neutral. If we had mechanisms accessible to moral reasoning that tracked the value of premarital sex independent of prevailing views in particular contexts about premarital sex, then it would be reasonable to expect the group whose prevalent views agreed with the actual value of premarital sex to give more unified responses than the group whose views disagreed with its actual value. However, so far as I can tell, this was not the case.

This line of thought seems to take a realist position on moral value, insofar as it treats it as a property of objects, states of affairs, etc. to which humans might have epistemic access. I do not mean to beg any meta-ethical questions here. I take the realist understanding of moral value to be the most natural understanding of the hypothesis of Knobe and Roedder. Moreover, consider the most direct alternative: If we are to avoid the realist construal of values in the second case, then we seem to be committed to the Manhattan parkgoers' constituting a group by whose perspective refraining from premarital sex has one valence, while from the perspective of the participants in the Mormon Bible-study group it has a different value. I take this to be a familiar, dubious form of "cultural" relativism, and hence not attractive as a construal of the nature of value.

Clearly, more work must be done before it will be tenable to make more confident claims about the mechanisms responsible for patterns found by experimental means in moral reasoning. The reflections about realism and anti-realism suggest that, at some points at least, experimental philosophy encounters limitations to be addressed by more standard philosophical methods. That is, if there are other meta-ethical positions that might make a cognitive difference, then tests could be designed to evaluate them. Perhaps something like R. M. Hare's prescriptivism or Simon Blackburn's quasi-realism provides a plausible position between outright realism and crude cultural relativism. Ultimately I hope that the discussions of moral dumbfounding and premarital sex help to show the empirical merits of the moral reasoning system hypothesis developed earlier in the chapter.

4 Rethinking the Reactive Attitudes: Attributing Moral Responsibility

This chapter is about the capacities required to attribute moral responsibility to ourselves and others. Suppose that someone does something nice for you. Flushed with gratitude, you thank the person for her benevolence. What are the cognitive capacities required for you to praise someone in this manner? Suppose that you steal from your neighbor. Your neighbor discovers this and flies into a rage, inveighing against you as a bad person. What cognitive capacities underlie such condemnation? This topic should be familiar, but it is overlooked in recent moral psychology. It is familiar both because of the regularity with which we attribute moral responsibility and because of some classic philosophical work on the subject.

This chapter focuses on P. F. Strawson's account of the attribution of moral responsibility. Strawson's 1962 essay "Freedom and Resentment" provides a starting point for many discussions of moral responsibility.[1] Nevertheless, this topic is rarely treated as an important part of an account of our core moral-psychological capacities. It barely appears in recent empirically informed books about moral psychology. John Doris addresses it a bit in chapter 7 of his 2002 book *Lack of Character*, but mainly in terms of what it is to *be* responsible. That is, Doris does not address the capacities by which we attribute moral responsibility to each other. Experimental philosophers have taken a look at this subject in connection with the concepts of freedom and responsibility. Perhaps this relative neglect is understandable. We reason about attributions of moral responsibility, so it may be reasonable to think that studies of moral reasoning subsume moral responsibility. It is easy to assume that seeing someone as morally responsible is no different, psychologically, from seeing someone as a human, or a person, or a neighbor, or a baker. That is, it is easy to assume that seeing people as morally responsible is only one way of characterizing people

among many other psychologically equivalent ways, and that this particular way of seeing people may be captured by studies of moral judgment. However, Strawson's work gives us reason to question such subsumption of moral responsibility under other topics. Strawson draws our attention to the psychological richness of attributions of moral responsibility. He presents the attribution of moral responsibility as drawing on affective capacities and on capacities for interpersonal interaction. Its psychological richness makes this topic a good one for direct treatment within empirically minded examinations of our moral psychology.

Understanding this psychologically rich territory cannot be done by philosophy alone; it requires the empirical resources of psychology. This topic has not yet received a definitive interdisciplinary treatment. I will not give it one in this chapter. However, I will draw on recent work in both psychology and empirically minded philosophy in pursuit of more modest aims, which are to sharpen our view of the psychological capacities required by our practices of attributing moral responsibility and to develop a hypothesis about the extent to which these capacities are realized in wide psychological systems. This discussion takes us through deep waters via a circuitous route, but this is apt for doing justice to the psychological richness of moral responsibility.

4.1 The Wide Moral Systems Hypothesis and Attributions of Moral Responsibility

Here is a look at the structure of the discussion that will follow and at the view I will eventually defend.

My foil in this chapter is a reconstruction of Strawson's position that I call the Tempting View. Those who take that position hold that having feelings of certain kinds is both necessary and sufficient to attribute moral responsibility. The feelings in question are located within the physical boundaries of individual agents. I will examine the sufficiency claim and the necessity claim of the Tempting View, and will find both claims wanting. In place of the Tempting View, I will offer a view of attributions of moral responsibility with four tenets:

The feelings emphasized by Strawson are indeed important to attributions of responsibility, but having such feelings is neither necessary nor sufficient to make such attributions

We attribute moral responsibility in various ways, using diverse psychological capacities.

Some of these capacities are locationally wide.

The psychology of attributions of moral responsibility has a thin unifying thread in our mind-reading capacities. Mind reading is necessary for attributions of moral responsibility. This is also psychologically heterogeneous and, to some extent, realized in locationally wide systems.

Let me begin with a brief description of Strawson's account of the attribution of moral responsibility.

4.2 Strawson: Moral Responsibility and the Reactive Attitudes

It is useful to divide Strawson's position on moral responsibility into two parts: (A) an account of our practices of attributing moral responsibility in terms of what he calls the "reactive attitudes" and (B) an argument that the account in part A shows us how moral responsibility and determinism could be compatible.[2]

Strawson's topic, common to many such discussions, is whether moral responsibility is compatible with determinism. Determinism is part of a thoroughly objective account of the world. To approach phenomena deterministically is to adopt a thoroughly objective attitude toward them. Strawson's strategy is to encourage the reader to think of as many kinds of interpersonal relationships as possible, then to think of the kinds of importance we attach to the attitudes and intentions directed toward us by the others in these relationships, and then to think of our own "reactive attitudes" (1962, p. 6 in 1974 reprint)—that is, our own attitudes of response to the attitudes and intentions of others. These reactive attitudes include personal ones, such as resentment, and more general ones that Strawson thinks include characteristically moral ones. Here is a list of examples from page 15 of Christian Smith's 2003 book *Moral, Believing Animals*:

> A son feels guilt for not taking care of his ailing, aged mother in a way he knows a good son should. A wife feels annoyed that her husband spends the weekend watching sports on television when he could be painting the house or talking with her about her week. . . . An employee feels angry for not getting the raise she thought the boss had promised and that she clearly deserves. The party host feels embarrassed in front of their guests by the rude misbehavior shown by their teenage kids. A clique of university students is elated to hear that the professor they had for a

class who was a terrible lecturer and an unfair grader was denied tenure. . . . A father feels profound contentment when his daughter eagerly takes over the family business that he started and built up over the last thirty-five years. A girl feels betrayed when she learns that the boy she was dating "exclusively" has also been seeing another girl. Passersby feel indifferent to a homeless beggar they suspect is a self-destructive drug addict. A group is offended when leaders of another religion organize to proselytize its members. A nation's people is shocked by the unprovoked attack of a neighbouring country and rallies to prepare for war.

Once we have listed the appropriate relationships and attitudes, the question to ask is whether the acceptance of the thesis of determinism could lead us always to look on everyone exclusively with the objective attitude (Strawson 1962, p. 11 in 1974 reprint). Doing so would mean giving up the subjective engagement of which we have reflectively framed an account. An affirmative answer delivers the incompatibility of moral responsibility with determinism. Strawson's answer, however, for both the personal and more general reactive attitudes, is that this is practically though not logically inconceivable.

There is a second aspect to Strawson's conclusion. On occasion we do adopt the objective attitude toward others. For example, when we find out that someone is mentally incapacitated in specific ways, we suspend our attitudes of resentment, and we cease to deploy the apparatus of moral responsibility in connection to the conduct of that person. The way we suspend our subjective engagement with such apparent amoralists is very important. Strawson contends that we do not take up the objective attitude as a result of conviction of the truth of determinism. Instead, the adoption of such an attitude is a consequence of the giving up of our subjectively engaged perspective. Determinism, or particular applications of this thesis, is never the cause of our suspension of our normal attitude. Further, we abandon our normal perspective for specific reasons in specific cases. We do not give it up wholesale as a result of a general theoretical conviction.

Overall, the lesson is that once we pay proper attention to the interpersonal domain in which the practices of moral responsibility have their home, it will become clear that determinism is no threat to moral responsibility. Morality and determinism are compatible because the objective domain that is the appropriate home of the discourse of determinism is, in a certain sense, irrelevant to morality. Strawson puts it this way:

[Q]uestions of justification are internal to the structure [of human attitudes and feelings] or relate to modifications internal to it. The existence of the general framework of attitudes itself is something we are given with the fact of human society. As a whole, it neither calls for, nor permits, an external 'rational' justification. (p. 23 in 1974 reprint)

4.3 A Tempting Interpretation of Strawson

Strawson's account of the attribution of moral responsibility is partly what we might call a "performance-level" description, in that it describes a distinctive domain of human activity. Insofar as Strawson also offers hypotheses about the psychological mechanisms by which this activity is produced, it is also a "psychological-level" description. A tempting and broadly Strawsonian hypothesis about the psychology of responsibility attribution can be generated by focusing our attention on specific elements in both the performance-level description and the psychological-level description.

First, consider what Strawson offers as the object or target of the reactive attitudes: the attitudes and intentions of others. Here are two examples:

My resentment of you could be triggered by and directed toward your malevolent intentions toward me or toward something about which I care.

My gratitude toward you could be triggered by and directed toward your benevolent intentions toward me or toward something about which I care.

This aspect of Strawson's account operates at both the performance level and the psychological level. At the performance level, it is at least a partial description of at least the target of attributions of moral responsibility—certain sorts of thoughts. This entails something at the psychological level: the mechanisms by which we attribute moral responsibility to each other include the mechanisms by which we understand the thoughts of others. The capacity to understand the thoughts of others is commonly characterized as "mind reading." On the broadly Strawsonian account being constructed here, the attribution of moral responsibility involves, perhaps quite centrally, our mind-reading capacities.

Second, consider Strawson's central psychological claim: that we attribute responsibility via the reactive attitudes. Here are some specific examples of reactive attitudes from Strawson's discussion: resentment, gratitude, forgiveness, indignance, disapproval, feeling bound, feeling compunction,

guilt, shame. Strawson also acknowledges a wide array of other reactive attitudes. These are all affective states. That is, the broadly Strawsonian account of the attribution of moral responsibility appears to gives central place to emotions of a particular kind. R. J. Wallace (1994) locates Strawson's account in a broader tradition of thought about the so-called moral sentiments.

If we combine these two observations with the distinctions between kinds of internalism and externalism that I introduced in chapter 1 of this book, we get a tempting interpretation of the Strawsonian account of the attribution of moral responsibility:

The Tempting View Feelings of a certain kind are both necessary and sufficient for the attribution of moral responsibility.

Since these feelings are directed at the thoughts of others, they are taxonomically wide. For example, gratitude in general, and certainly the specific instances of gratitude that fall under Strawson's account, must be understood as a particular kind of feeling toward the attitudes and intentions of others about oneself as revealed in the actions and utterances of those others. However, these feelings are locationally narrow: they occur strictly within the physical boundaries of the agent who experiences them. The various ways by which we might express these feelings—including actions and utterances—are constitutively distinct from the feelings themselves, and hence from the psychological mechanisms by which we attribute moral responsibility.

4.4 Emotional Perception

The Tempting View can draw some support from recent work on the nature of emotions. In *Gut Reactions* (2004), Prinz defends an account of emotions that casts them as a perceptual capacity. Specifically, Prinz argues that emotions are perceptions of bodily changes and, via these, of "core relational themes" (224–225). Core relational themes are, roughly, relations an individual has to his or her environment that pertain to that individual's welfare (15–16). In general, this is an attractive way of understanding something of what is involved in experiencing the reactive attitudes. When I feel indignance at a personal slight, perhaps it is apt to interpret this as my perceiving something about another person and that person's thoughts

and conduct toward me. Just as, say, the experience of different qualia is a way of registering differences in the reflective capacities of surfaces, perhaps the experience of the various distinct feelings characteristic of the reactive attitudes is a way of registering differences in how the thoughts and behaviors of others can matter to me. Indeed, Prinz uses some of the reactive attitudes as examples. For example, he describes shame as "a sense of unwelcome attention that occurs when one has committed a transgression that will disappoint others" (156) Prinz discusses guilt in connection with other topics, and characterizes it as a kind of sadness at one's own transgressions (124–129).

Besides providing an attractive account of the nature of the reactive attitudes, Prinz's position provides a line of support for the Tempting View via the overall picture of the structure of the mind in which he locates his more specific account of emotions. A natural way to study emotions as a variety of perception is to use other perceptual capacities as a model: identify important characteristics of uncontroversial perceptual capacities, then see whether emotions have identical or similar features. This is Prinz's method on pages 221–222 of *Gut Reactions*, and throughout that chapter. One of the characteristic features of ordinary perceptual capacities, and the only one which will concern us here, is their modularity.

Following Prinz (2004, 232), here is a list of the hallmarks of classical modularity. (This is a standard list; see also Fodor 1983 and Karmiloff-Smith 1992.) Modules are

localized in dedicated parts of the brain,

subject to characteristic patterns of breakdown,

automatically operating,

rapidly processing,

productive of simple outputs,

inaccessible, in that their inner workings are relatively closed to higher levels of processing,

informationally encapsulated, in that their processing cannot be guided by information at higher levels of processing,

ontogenetically determined,

and

operative over a restricted and specific domain of inputs.

Prinz argues (2004, 232–236) that emotions share these characteristics (more or less; both he and Fodor acknowledge that modularity can be a matter of degree).[3] This means that they share a very important feature with uncontroversially perceptual capacities, such as vision. Thus an important part of the overall case for emotions as perceptual capacities is in place.

This view of perceptual modularity is connected to a particular overall view of the structure of the mind. Prinz gives us an explicit statement of this view:

> The mind is divided into different kinds of information-processing systems. There are perceptual systems that provide inputs, action and motor systems that provide output, and, perhaps, higher cognitive systems that engage in reasoning, planning, problem solving, and other mental operations that mediate between inputs and outputs when we move above the level of reflex response. (2004, 221)

If emotions are perceptual capacities, then they fall into the input part of this view of the mind. In all likelihood, they provide information to both the action-production systems and higher cognition. On this view, both higher cognitive processes and action-production systems are constitutively distinct from all perceptual capacities, including emotion.

Let's connect this to the Tempting View. This interpretation of Strawson holds that the reactive attitudes are necessary and sufficient for attributions of responsibility. This entails that various means of expressing the reactive attitudes, and thereby our attributions of responsibility, are all constitutively distinct from the attitudes themselves. This is exactly the view Prinz takes of emotions in general: as a perceptual input faculty, they are constitutively distinct from both higher cognition and the various sorts of output by which we might express our emotions. Since it is plausible to see the reactive attitudes as affective states, Prinz's account of the nature of emotions seems to provide a natural home for the Tempting View. To the extent that Strawson is right about responsibility and that Prinz is right about emotions, the Tempting View is exactly what we should expect.

Despite all this, there are reasons to think that the Tempting View must be modified. I shall argue that it must be modified to such an extent that we might as well see it as mistaken. The modifications are required at both the performance level and the psychological level. The Tempting View's claim that certain locationally narrow feelings are necessary and sufficient for the attribution of moral responsibility must be rejected. The eventual

account of moral responsibility that will be offered in place of the tempting interpretation of Strawson's position reveals our capacities for such attributions to be locationally wide to a significant degree.

4.5 The Sufficiency Claim of the Tempting View: Vertical and Horizontal Modularity

I shall begin by examining reasons to question the sufficiency claim of the Tempting View, which is that having feelings of a certain kind suffices to attribute moral responsibility. I will do this by focusing on the overall picture of the mind in which the Tempting View finds its natural home. Susan Hurley has called this view of the mind into question. She calls it the "classical sandwich" view of the mind (1998, 20–21): higher cognition is the filling between the input and output layers. The input and output layers have more contact with the environment. (See figure 4.1.)

Hurley describes this view of the mind as being structured with *vertical* modules: specific perceptual capacities constitute the leftmost vertical layer of figure 4.1, and they are constitutively distinct from the processes that constitute the other vertical layers. Fodorian modules belong to this vertically modular, classical sandwich view of the mind. If emotion is a classically modular perceptual capacity, then it belongs in the leftmost vertical column.

Hurley argues that this view of the mind, and specifically its vertical modularity, has been called into question by neuroscience. Instead of a

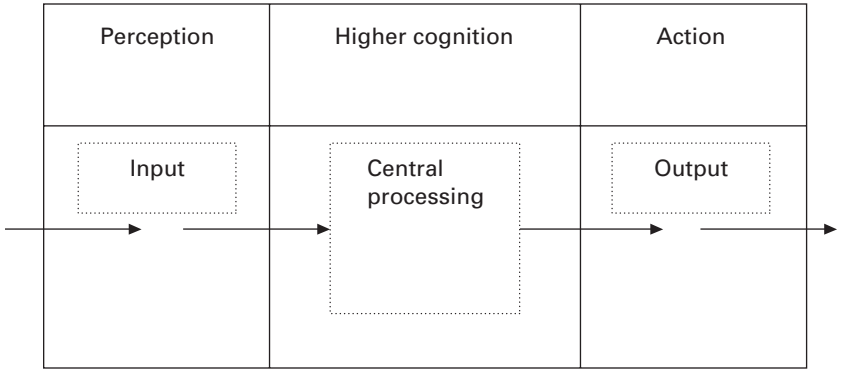


Figure 4.1

mind composed of constitutively distinct vertical modules, Hurley argues, neuroscience reveals a mind structured by *horizontal* modules. Horizontal modules are content-specific and task-specific systems that "[loop] dynamically through internal sensory and motor processes as well as through the environment" (1998, 21; 408). They are "modular" in virtue of their content-specific and task-specific functionality. First, each module is constituted by *both* input and output functions. Functioning *within* each module can include feedback from relatively more downstream stages of processing to relatively more upstream stages. Second, there is no modular layer that, by itself, constitutes higher cognitive functioning. Instead, this is something that emerges from the interplay of the specific perception-action layers. (See figure 4.2. This view of the mind is, of course, not limited to two horizontal layers. The broken lines indicate porous boundaries between input, higher cognition, and output, in contrast to the more rigid vertical

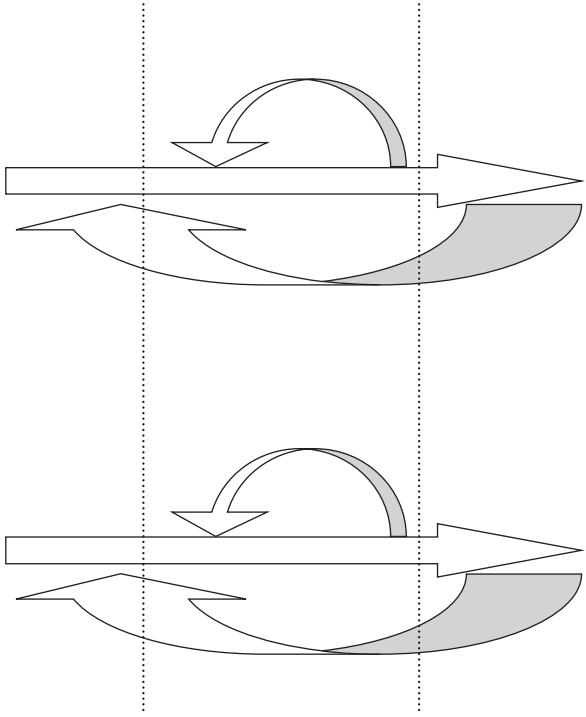


Figure 4.2
Horizontal modularity. Adapted from p. 407 of Hurley 1998.

separations of the classical sandwich view. Moreover, there can be complex relations between horizontal modules, but it is not necessary to represent them in the diagram for present purposes.)

If Hurley is correct that recent findings from neuroscience point toward a horizontally modular mind rather than a classically vertically modular one, then the task of modeling emotional perception on other perceptual capacities is complicated. In one sense, the task of describing the features of perceptual capacities has to be done more in a case-by-case manner than it had seemed before: maybe some modalities are classically modular, whereas others are horizontally modular. However, in another sense, just a little horizontal modularity seems to pose a large problem for vertically modular views of perception. Horizontal modularity of any kind seems to call into question the overall view of the mind in which vertical modularity has its natural home. Hurley claims that it challenges the status of the classical sandwich as a general conceptual framework for thinking about the mind. With regard to emotion research, Hurley's suggestion complicates the matter of assessing such views of emotional perception as the one offered by Prinz. How then should we proceed?

Although vertically and horizontally modular accounts of perception and the mind share certain features (e.g., Hurley thinks horizontal modules are domain specific), they differ in ways that allow for empirical testing in general, and for empirical testing of emotional perception in particular. Such testing might provide direct evidence for one of these views of the modularity of emotional perception. Given that such testing has been done only indirectly, we must be careful about the conclusions we draw. Nevertheless, the territory as it is suggests something different from the view of the mind that grounds the Tempting View.

Before we look at some possible avenues of empirical assessment, more attention must be given to vertically and horizontally modular models of emotional perception. The resulting refinements will give us a better view both of what might be empirically tested and of how such testing could be conducted.

4.6 Refinements

The focal point of this discussion is the views of modularity associated with two different accounts of perception in general, and of emotional

perception in particular. That is, the topic is the *structure* of emotions, and in particular of the kinds of processing that realize emotions. I take it that a particularly important way to study this topic is to study the neural processes that implement such processing. Hence, the refinements I shall attend to have to do with neurobiology and vertical and horizontal modularity.

Vertical Modularity

In a consideration of possible objections to the view that emotions have a classically modular structure, Prinz distinguishes two different sorts of pathways that constitute emotional processing (2004, 234–236). I presented this distinction in chapter 2, but since the details are important to the present discussion I repeat them here.

First, there are *initiation* pathways. These may be thought of as the input routes to the emotional module. Their general job is to receive input from a variety of sources and then to prepare this input in a manner appropriate to the remainder of the emotional processing. As an example, Prinz discusses the role of the amygdala in the processing of fear, disgust, and sadness: "The amygdala receives inputs from a variety of different brain regions and initiates a pattern of bodily outputs, which then give rise to these emotions." (2004, 234) An important feature of the initiation pathway is what Prinz calls *calibration files*. Calibration files are sets of representations linked to particular bodily responses. Prinz holds that such files allow us to modify emotions via judgments. The establishment of new calibration files allows us to modify emotions (more specifically, embodied appraisals) to apply to things other than those to which they evolved to apply (99–100).

Second, there are *emotion response* pathways. Crucially, Prinz holds that this is where we find emotions themselves. Strictly, on this view the initiation pathways are constitutively distinct from the modules of emotional perception. This means that the features of classical modularity apply to the response pathways alone. Take domain specificity as an example: Prinz claims that the amygdala is not domain specific, but that the response pathways that receive information from the amygdala about objects of fear, disgust, or sadness are (234).

Accordingly, emotional processing has the structure illustrated in figure 4.3.

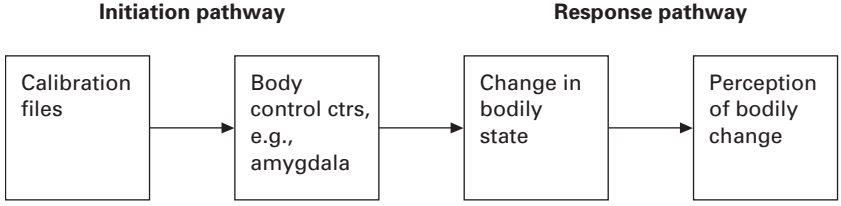


Figure 4.3
Emotional processing. Adapted from p. 235 of Prinz 2004.

Although it is important to distinguish initiation and response pathways, we have reason to question the details of this picture of emotional processing. In particular, recent research calls into question the confinement of the amygdala to input preparation. Richard Davidson and William Irwin (1999, 15) claim that the amygdala is important for both the perception and the production of negative emotion. In a recent summary, Elizabeth Phelps claims that research shows that the amygdala has a critical role in both the acquisition and expression of fear learning (2004, 1005 and throughout). Glenn Schafe and Joseph LeDoux give the amygdala a central role in both input and output pathways in fear conditioning (2004, 987–989). Figure 4.4 is a simplified version of their model for conditioning to an auditory stimulus.

In particular, Schafe and LeDoux claim that the relations between the various parts of the amygdala are essential for fear expression (2004, 989). Even more important are the studies of emotion regulation that I discuss later in this chapter. Overall, the working assumption in studies of the neuroanatomy of fear seems not to be that the amygdala is solely a component of an input pathway to a fear module, but that it could be more intrinsically woven into the processes that realize fear itself. Even if all of this does not constitute conclusive support for inclusion of the amygdala in the response pathway of fear, I take it to show that the evidence from neuroscience does not currently support a rigid boundary between the functioning of the amygdala and, in this case, the neural processes that realize fear.

Horizontal Modularity

Hurley defends a position that she describes in terms of two-level interdependence. Recall the diagram of the classical sandwich view of the mind.

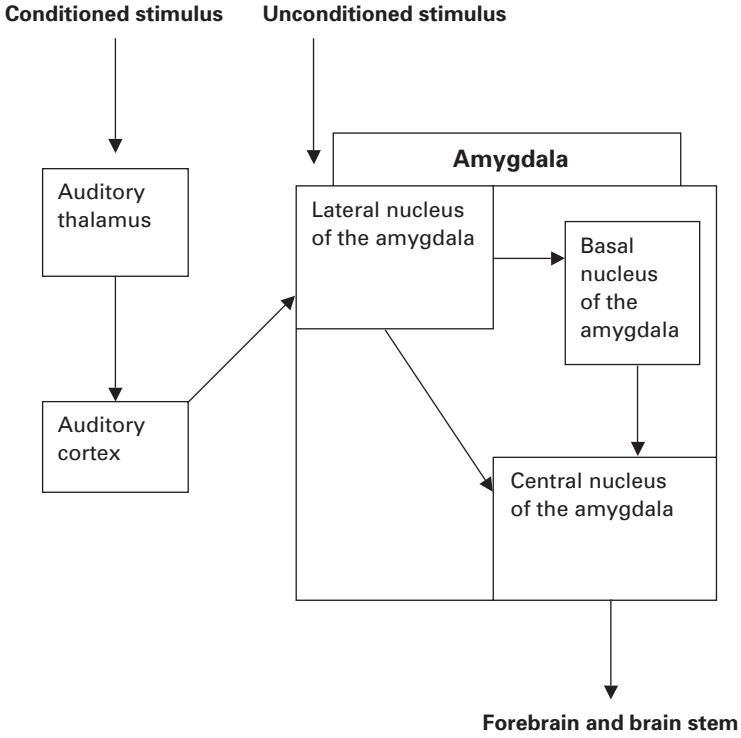


Figure 4.4

The two levels in question are the personal and the subpersonal. The personal level is characterized in terms of perceptions and actions—things a person might experience or do. The subpersonal level is described in terms of input and output processes—that is, in terms that are appropriate to discussions of mechanisms rather than persons. One of Hurley's principal concerns is that the classical sandwich picture of the mind simplistically maps personal and subpersonal levels onto each other. Hence, distinctions found at the personal level that are described in terms of perceptual content and the content of intentions respectively are taken to be functions of distinctions in input and output respectively. Hurley argues in a variety of ways that both distinctions and invariants in personal-level content can be functions of *relations between* input and output. In such cases there is no one-to-one mapping of personal-level distinctions to subpersonal-level distinctions. Hence, the alternative view of the mind has,

so far, been described here in terms of horizontal modules constituted by feedback relations that cut across the subpersonal boundaries between input and output. And higher cognitive processes have been alluded to as emerging from the interplay of horizontal modules, rather than depending on central subpersonal processes distinct from both input and output systems.

However, another of Hurley's themes modifies this stark contrast with the classical sandwich view of the mind. This is her commitment to empirical study of the processes that realize mentality. Accordingly, she thinks that it is an empirical issue whether a given kind of mental processing is realized by classical, vertical modules or by horizontal ones with both input and output functioning. Interestingly, as a result of this commitment Hurley seems to be committed to the existence of more and less centralized processes. For instance, much of her case depends on arguing that certain studies in neuroscience show, e.g., the dependence of perceptual content on output systems. She considers an objection to her case that claims that perceptual content depends on input and central processing but not on output (1998, 383–384). If Hurley were really committed to there only being input and output systems, then she could dismiss this objection as relying on a class of neural processes that do not exist. But this is not her reply; instead, she acknowledges the possibility of centralized processing, but denies that the line of thought in the objection can deliver a principled distinction across sensory and action modalities between central and peripheral processing. She also thinks that such an objection already admits her point, which is that personal-level content can depend on relations between different sorts of subpersonal processing rather than simply mapping onto a single kind of subpersonal processing.

All of this calls for a refinement of our view of horizontal modularity. Instead of there being only one kind, now there are two broad kinds to consider: full and abbreviated horizontal modularity.

Full horizontal modularity is horizontal modularity as originally described: domain specific modules constituted by both input and output processing, and hence cutting across the full width of the mind.

Abbreviated horizontal modularity is horizontal modularity that does not cut across the full width of the mind from input to output, but instead cuts across some of the vertical boundaries that are characteristic of the classical sandwich view of the mind. Given the tripartite division of the

classical sandwich view, there are two possible kinds of abbreviated horizontal modules: input–central processing modules (in which relations between input processing and more centralized processing subserve some specific sort of personal-level perceptual content[4]) and central processing–output modules (in which relations between more central processing and output processing subserve some specific sort of personal-level intentional content).

The crucial general feature of horizontal modularity is that relations between different kinds of neural processing get a constitutive role with regard to the individuation of the modules subserving personal-level content. On the classical sandwich view, relations between different sorts of processing can have an instrumental role only, never a constitutive role. For present purposes, discoveries of both full horizontal modularity and input–central processing abbreviated horizontal modularity for the reactive attitudes would undermine the Tempting View. Discovering that these were reasonable ways to see the mind in general, without specifically discovering this for the reactive attitudes, would present the Tempting View with a challenge to its *prima facie* plausibility.

4.7 Empirical Suggestions

Given the differences in structure between vertical and horizontal modules, we should be able to design tests to determine which view, if either, fits perception in general and specific sorts of emotional perception in particular. I have three suggestions—two rather sketchy ones and a more developed one—for such testing.

The Role of Inhibition

This is the first of the sketchy suggestions. One thing that might be examined is the role of inhibition in the two models of perceptual modularity. The issue I have in mind is not inhibition of personal-level behavior. For instance, some models of certain psychopathologies posit inhibition systems. It is the failure of these systems in certain ways that is thought to give rise to the psychopathologies in question. (For a discussion of a similar approach, see Kring and Bachorowski 1999.) Instead, what should be examined to adjudicate between vertical and horizontal modularity is inhibitory factors in information processing. As I see it, each of the two

models has a role for inhibition in normal and abnormal processing, but they differ in the details. In particular, it seems that they will differ in the extent to which they require inhibition of information processing for normal functioning.

On the vertical model, it is natural to think of inhibition as primarily constituting an obstacle to normal functioning. Here is a description of this sort of modular processing:

> According to Fodor, information from the external environment passes first through a system of sensory transducers, which transform the data into formats that each special-purpose input system can process. Each input system, in turn, outputs data in a common format suitable for central, domain-general processing. (Karmiloff-Smith 1992, 2)

Perhaps there is inhibition of processing of various kinds *within* the various stages of this linear process, but the overall image is of an assembly line: input is processed in some way, then passed along to the next stage for a different sort of treatment. This picture invites the thought that if there is inhibition of this process, it will most often constitute a problem rather than facilitate normal subpersonal processing of information and normal personal-level experience and behavior.

In contrast, on the horizontal model, inhibition of the work of particular modules is central to normal subpersonal and personal-level functioning. Hurley explicitly (though tentatively and speculatively) gives a role to the inhibition of the deliverances of horizontal modules in the production of rational action (1998, 409–412). She has in mind fully horizontal modules: ones that are constituted by both input and output processing. She argues that horizontal modules are implicated in the patterns of imitation exhibited by newborns, normal adults, and patients suffering from certain sorts of brain damage: input processing classifies the behavior of others as of a certain kind, and output processing produces behavior of the same kind. Imitation has beneficial effects, such as providing developmental ways of calibrating motor systems and acquiring basic intentional capabilities (ibid., 411). But insofar as imitation verges on reflex behavior, Hurley argues that it brings with it a threat to rationality: "Imitation need not be merely reflexive, but can entrap cognitive processes. This is typical of a horizontal module, considered in isolation from others." (410) Rational action is achieved not through isolated horizontal modular functioning but from the interaction of multiple layers of such modules. Crucially,

such layered functioning involves inhibition of the functioning of some modules. "Rationality can be conceived as an emergent property of such a complex system. . . . Rationality may emerge from complex relationships between horizontally modular subpersonal systems which, considered in isolation, generate behavior that is less than rational." (412) If modules interacting in this way could be found for emotions, then *prima facie* they would be psychologically consistent with rationality: inhibition would be the means by which the threats to rationality posed by fully horizontal emotional modules could be handled.

Besides the general topic, there is a particular implication of this issue for traditional interests of philosophers. For Hurley, horizontal modules are automatically motivating. An absence of motivational effect is an achievement of a system of such modules. In contrast, Prinz denies that emotions are always motivating. Instead, he says they always provide "motives" rather than "motivations." Motives are reasons for action, whereas motivations are psychological impulses that actually produce behavior (2004, 193). On the classical, vertical view of the mind, rationality and rational behavior are a matter of how central processing utilizes the data from the input modules. The contrast between the views is this: by the standards of the vertical model, the deliverances of emotional modules are (primarily) information for rational consideration. This information will deliver behavior *only if* central processing *converts* it into a motivation. The flow of information is linear, from input to output via central cognition. By the standards of a fully horizontal model, the deliverances of emotional modules are *automatically* motivators and secondarily sources of information for the processes that realize rationality. Rational thought and behavior emerge from the inhibition of the processing of such modules.

These vague remarks suggest the following avenues of more specific research that may shed light on the kind of modularity characteristic of emotional perception:

Development. Hurley's remarks about imitation and newborns point to a general domain of research that ought to be rich in data about the role of inhibition in normal and abnormal processing. Hurley's suggestion for imitation is that maturation is marked by increased inhibition. In contrast, Annette Karmiloff-Smith's view of development (1992) is of a process of increasing vertical modularity. This would not necessarily bring the sort of inhibition of processing that Hurley speculates about. With regard to

emotion, what is needed is a comparison of the neural processes of adults and infants. If there is some way of characterizing these processes in terms of more and less inhibition, then there should be a way of determining whether emotional neural development through normal maturation is characterized more by vertical modularity or by horizontal modularity.

Psychopathology. It is common for psychopathologies to be characterized by emotional problems. A natural general hypothesis, though one that must be refined in many ways, is that such emotional problems result from obstacles to the neural processes responsible for emotional experience and related behavior. For example, psychopaths are well known to have emotional deficiencies (Hare 1993; Prinz 2004; Blair et al. 2005). Given the apparent differences in the role of inhibition in emotional processing in the vertical and horizontal models, and given the hypothesis that many emotional problems are due to inhibition of normal emotional processing, specific applications of the horizontal and vertical models ought to predict different kinds of inhibition as the cause of particular psychopathologies. However, a caveat is warranted here: If an emotional problem is due to abnormal development, then we cannot assume that it is constituted by inhibition of otherwise normal neural pathways. This is probably the case with psychopaths, since the signs of psychopathy show up early in life. Thus, the primary relevant sort of psychopathology for adjudicating between vertical and horizontal modularity is the sort that arises in the absence of abnormal emotional development.

Empathic Emotional Recognition

This is the second of the sketchy suggestions. An important part of Hurley's case concerns neurons and neuron populations that are important to both perception and action. So-called mirror neurons have been discovered both in monkeys and in humans—Prinz discusses them briefly (2004, 229); for lengthier discussions, see Rizzolatti et al. 1996; Keysers et al. 2003; Gallese et al. 2004. These neurons are activated both when a monkey (or a human) observes another monkey perform an action of a particular kind and when a monkey performs the same kind of action itself.

Can similar cases be found for emotion? The sort of phenomenon to look for is *perception* of emotion and related behavior in others that is processed with at least some of the same neural circuitry as the *production*

of the same emotion or emotional behavior. A promising starting point is Robert Gordon's model of the general psychological capacity of empathy. Broadly taken, empathy involves sharing the feelings of others. When I feel what you feel, I am empathizing with you. This is also referred to as "emotional contagion"—feelings are catching, in that exposure of one person to another person displaying certain feelings often produces the same feelings in the observing person. However, it is reasonable to differentiate between different sorts of empathy. (For recent psychological work that attends to different sorts of empathy, see de Vignemont and Singer 2006 and especially Singer 2006.) For present purposes, full empathy or emotional contagion is not the issue. Instead, the relevant phenomenon to search for is what I shall call "empathic emotional recognition": perception of emotions in others, without experiencing those emotions oneself, that involves the same neural basis as the experience of those emotions. The role of the same neuron populations in both the production and recognition of emotions is what makes this broadly "empathic." Gordon's work is a promising starting point because he models empathy in terms of simulation. In general, simulation consists in "off-line" use of one's own cognitive apparatus to take the perspective of another person. "Off-line processing" means that the cognitive processes in question are detached from the normal routes leading to practical decision making and action for oneself. One is using them not for oneself but rather to understand the perspective of another, so they need not be plugged into these practical processes. Simulation of the feelings of another would require the use of the neural circuitry that realizes the feelings in question for oneself, but detached from normal processes delivering emotional experience and related conduct (Gordon 1995).

Here are two more detailed versions of the kind of empirical test suggested by this line of thought:

(i) Specify, even roughly, the neural processing of a certain emotion. With this information in hand, see whether the same neural circuitry is implicated in recognition of this emotion in other people.

(ii) Specify, even roughly, the neural processing of a certain emotion *plus* some sort of typical response. With this information in hand, see whether the same neural circuitry is implicated in recognition of this emotion *plus* response in other people.

The first test requires that we are able to recognize emotions in other people detached at least from typical responses, and maybe from all expressions. This may be impossible; I leave this to be determined *a posteriori*. The second test does not require this. If the results for (i) vindicated the simulation model of empathic emotional recognition, then we would have empirical support for abbreviated horizontal modularity for the emotion in question. This kind of test supports abbreviated horizontal modularity because action is not implicated in what is recognized. All that would have been discovered was that the same neural circuitry was involved in experiencing and perceiving a particular emotion. In contrast, (ii) would provide evidence of full horizontal modularity. The reason is that this test would show that the same neural circuitry is implicated in both the perception of the combination of a particular emotion and an associated response and in the production of the combination of the first-person experience of the same emotion and the first-person performance of the same response.

Fear provides a promising test case, since there are already well-developed methodologies for studying both the experience of fear and the expression of fear responses. For example, recognition of objects or states of affairs that are threatening, and that hence call for a fear response, is accompanied by a distinctive eye movement known as the startle eyeblink. Studies of fear have combined subjective reports of emotion, measurements of the startle eyeblink, and fMRI scanning. (See Ochsner et al. 2002 and Schaefer et al. 2002, and the references in these papers, for work utilizing this combination, as well as other approaches.) Accordingly, here is a first pass at modifying this methodology to study the form of modularity of fear: Subjects could be divided into two groups: a "fear group" (which would be subjected to fear stimuli) and an "observing group" (which would observe the first group). Specifically, the observing group would watch the faces of the fear group. Startle eyeblink could be used as an objective measure of fear response in the fear group.[5] It, and other facial expressions, would be the behavioral evidence of fear for the observing group. Functional magnetic resonance imaging would be performed on both groups. If the scans revealed significant overlap of neural processing in both the fear group and the observing group, we arguably would have evidence of full horizontal modularity of fear processing.

In fact, studies of the neural processing of fear and of recognition of facial expressions of fear in others suggest this sort of overlap of neural

processing. As I have already mentioned, the amygdala has been shown to be an important part of the neural foundations of fear learning and experience. PET and fMRI studies have also shown greater amygdala activity when people observe fear facial expressions than when they see facial expressions of other emotions. Davidson and Irwin (1999, 15), in a review of such studies, note that this goes even for faces that are not consciously noticed. This suggests that the perceiving subject is not actually experiencing fear as a result of perceiving the faces. This seems to amount to off-line activity of an important part of the neural basis of the experience of fear.[6] In other words, empathic emotional recognition seems to have been found for facial expressions of fear.

To find the general pattern of empathic emotional recognition for a particular emotion such as fear, it would be desirable to study an array of behaviors, not only facial expressions and certainly not only the startle eyeblink. It is not difficult to devise a wide array of stimuli for the observing group—actors and films are readily available. However, it is difficult to devise procedures to generate real fear and the particular desired fear behaviors, as opposed to the simulated ones that actors would produce. Consequently, this methodology might be confined to a fairly limited range of fear behavior, and hence the information it provides might be quite limited. Other emotions might be more easily operationalizable.

One such example is disgust. In a study of 14 subjects, Wicker et al. 2003 found empathic emotional recognition for real disgust in response to odors. Subjects each participated in four runs. In two runs, they watched movies in which people smelled the contents of a glass and responded either neutrally, disgustedly, or pleasurably. In the other two runs, subjects themselves experienced olfactory stimuli. These were either pleasant or disgusting. In all four runs, fMRI observations were made of subjects' neural activity. The most significant finding is that "the anterior insula is activated both during the observation of disgusted facial expressions and during the emotion of disgust evoked by unpleasant odorants" (Wicker et al. 2003, 655). Wicker et al. hold that this supports a "hot" theory of emotional recognition as opposed to a "cold" one. Cold theories accord emotion recognition roles to neural systems not directly involved in the experience of emotion. Hot theories hold that "brain areas responsible for experiencing [an] emotion will become active during the observation of that emotion in others" (655). This is not quite right, however. The

recognition of disgust and fear in others does not necessarily bring with it the experience of disgust or fear in the observer, even though such recognition involves neural activity that significantly overlaps with that found when the relevant emotion is experienced. This would be full-blown emotional contagion. Empathic emotional recognition without contagion is more directly analogous to action recognition: mirror neurons are activated both when someone observes a given action and when that person performs it, but recognition of the action is not necessarily accompanied by performance of it by the observer; far from it!

These studies are no more than suggestive with regard to the reactive attitudes. They provide us with reason to take seriously horizontally modular processing in both the theoretical modeling of these emotions and the devising of hypotheses to test these theories. They do not, however, show that that the reactive attitudes are horizontally modular. The design of studies for the reactive attitudes requires identifying important behavioral and facial expressions of these attitudes, then examining the neural processes responsible for recognizing these expressions in others and for producing these expressions oneself. Obvious candidates are blushing and the hanging of one's head associated with shame.

Non-Instrumental Content Dependence

Finally, here is a more fully developed empirical suggestion for assessing whether vertical or horizontal modularity characterizes emotions. The crucial issue is how the content of experience is related to the subpersonal processes which realize it. Return again to the diagram of the classical sandwich. On this view of the mind, personal-level content is mapped on to subpersonal-level processes in a one-to-one fashion. This limits the ways that subpersonal output processes can influence personal-level perceptual content. On the vertical view of perceptual modules, the only way output or central processing can affect perceptual content is *instrumentally*, i.e., by bringing about changes in input to perceptual mechanisms. The reason for this is that perception and action mechanisms, as well as central cognitive processing, are conceived of as constitutively distinct on this view. Such instrumental dependence is also possible on the horizontal view, but there is another possible kind of content dependence in this case. Since action and perception are constitutively interwoven on this view, it is possible for there to be *non-instrumental* dependence of perceptual content on output

processing on this view. This is revealed by changes in perceptual content even when input to perceptual mechanisms is held constant. Since such variation cannot be explained in terms of variation in input, feedback *within* the module from variation in other kinds of subpersonal processing to perception must be invoked instead. If non-instrumental content dependence can be found for emotional experience, then we would have evidence for horizontal modularity and against vertical modularity.

Hurley argues that evidence from studies from neurophysiology, marshaled in thought experiments based on such studies, shows that non-instrumental content dependence can be found for visual perception.[7] The present task is to devise studies that would do a similar sort of testing for emotional modules. The crucial thing to do is determine whether changes in emotional perceptual content can be brought about even when the input to emotional processing is held constant. If this is possible, then we have evidence that at least certain emotions are horizontally modular. If it turns out that this is not possible for certain emotions, then we have failed to find an important kind of empirical support for the horizontal modularity of these emotions.

For present purposes, abbreviated horizontal modularity will be our stalking horse. The reason is that much work has been done on the regulation of emotions. Broadly put, emotion regulation consists in the use of attention or some cognitive strategy to alter one's emotions (Phelps 2004, 1007). If extant work on emotional regulation shows that emotional perceptual content can depend on relatively more central processing without changes in input, then we have empirical support for abbreviated horizontal modularity of emotion. If this work does not show this itself but points toward ways of empirically testing whether emotional perceptual content can depend on relatively more central processing without changes in input, then it provides us with ways of empirically assessing the sort of modularity of emotional perception by assessing the possibility of non-instrumental emotional content dependence.

In one study (Schaefer et al. 2002), subjects were asked to view photos of two kinds: negative and neutral. Following exposure to the photos, there was a short delay period, after which subjects were asked how they felt. During both the viewing and delay periods, subjects were asked either to respond passively to the pictures—to let the emotional process that they triggered happen without any conscious intervention—or to maintain the

emotion that the photo triggered. Functional magnetic resonance imaging was used to see what was going on in subjects' brains while viewing the pictures, either responding passively or maintaining their emotion, and reporting their feelings. The results were as follows: Subjects asked to maintain their emotions reported stronger negative feelings in response to the negative pictures. The fMRI imaging revealed prolonged amygdala activity in maintain trials compared to the passive response trials (ibid., 913). Overall, this study provides evidence both that subjects can consciously regulate their emotions and that this is done, for negative emotions and at least in part, by affecting the activity of the amygdala.

In a similar study (Ochsner et al. 2002), subjects were shown negative pictures for a period of four seconds, followed by another four second period. During the second period, an instruction appeared in the viewing field. Subjects were given one of two instructions. The instruction *attend* required subjects to pay attention to the emotions triggered by the pictures without trying to change them. *Reappraise* required subjects to try to diminish negative emotions triggered by the pictures (ibid., 1217). Eyeblink startle tests had already confirmed that subjects could diminish their negative emotions through cognitive reappraisal: subjects reappraising their emotions had a smaller startle eyeblink magnitude than subjects not doing this (1216). In this subsequent part of the study, fMRI was used to discern the neural mechanisms of such reappraisal. In reappraise trials as compared to the attend trials, there was increased activity in the dorsal and ventral regions of the left lateral prefrontal cortex (LPFC) and the dorsal medial prefrontal cortex (MPFC); the LPFC appears to be more important. There was greater amygdala activity in the attend trials than in the reappraise trials (1220); there was also greater activity in medial orbitofrontal cortex (MOFC). Interestingly, there was a significant correlation between increased LPFC activity and decreased amygdala activity, and vice versa. This suggests that the neural mechanisms for regulation of negative emotions have, as important components, connections leading from the LPFC and MPFC to the amygdala and/or MOFC. Apparently, emotion is regulated via LPFC and MPFC suppression of activity in the amygdala and/or MOFC. Even further, connections between the LPFC and the amygdala appear to be particularly important.

What does all this mean for the project of assessing whether vertical or horizontal modularity is an apt model of the structure of emotional

perception? Well, these studies suggest that the content of emotional experience can be affected by relatively more central processes. More specifically, Schaefer et al. note that the initial activity of the amygdala is not affected by reception of the regulation instruction before exposure to the picture (2002, 915). It is later activity that the instruction affects. This is consistent with the role of feedback mechanisms in horizontal modules as Hurley describes them. What may be happening is that information initially processed by the amygdala is used by parts of the brain, such as the LPFC and MPFC, that realize regulation of emotion, such that subsequent amygdala activity is affected after feedback. This is speculative, but something like this is necessary for abbreviated horizontal modularity. More work needs to be done here.

Even so, showing these things to be the case would not suffice to provide empirical support for abbreviated horizontal modularity. For this, at least two things remain to be shown. The first is that the amygdala is part of the response pathway of the relevant emotions, not merely a part of the initiation pathway, as Prinz holds.[8] As I earlier mentioned, current work in neuroscience seems not to provide grounds to decide this one way or the other. If it turns out that the amygdala is relegated to the initiation pathways for these emotions, then we will not have evidence for abbreviated horizontal modularity. The reason is that it would seem that regulation of negative emotions was working only by affecting the input to the relevant processing. This would constitute *instrumental* dependence of emotional content on relatively more central processing. But *non-instrumental* content dependence is required for abbreviated horizontal modularity, and this requires the possibility of central processing affecting emotions without changing the input to these emotions. The second thing that remains to be shown is that, even if the amygdala is part of the response pathway for the emotions in question, the regulation of the relevant emotions is accomplished while input to the amygdala (or MOFC) is constant. These studies did give this issue some attention: Ochsner et al. instructed subjects reappraising not to look away, nor to distract themselves with extraneous thoughts (2002, 1225). Likewise, Schaefer et al. instructed subjects not to look away (2002, 918). However, this hardly seems adequate. If, as seems plausible, emotional processing can take thoughts as input, then instructions not to look away nor to think about extraneous things cannot guarantee constant input. More stringent

measures are needed to ensure that input to emotional processing really is constant while cognitive regulation of emotions is attempted. Two broad possibilities are (i) very close monitoring of input, to see whether there are significant differences between, e.g., attend and reappraise trials, and (ii) electrical or chemical regulation of input so that the experimental protocol controls its variation and constancy.

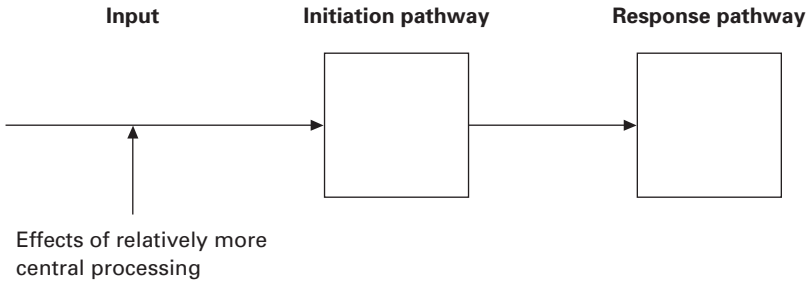
Even if input to, e.g., the amygdala can be held constant, subtler issues remain. Consider Prinz's division between initiation and response pathways. Calibration files constitute part of the initiation pathway. These files allow flexibility in patterns of emotional response which is determined by judgments—i.e., by cognitive control mechanisms. In the face of constant input to the initiation pathway, changes to the calibration files would facilitate changes to the information sent to the response pathway. Although input is, in such a case, constant in one sense, it is not constant in another. Accordingly, we can distinguish between two kinds of non-instrumental content dependence:

Weak non-instrumental content dependence Effects on content are brought about by changes to the calibration files without changes to the input to the initiation pathway.

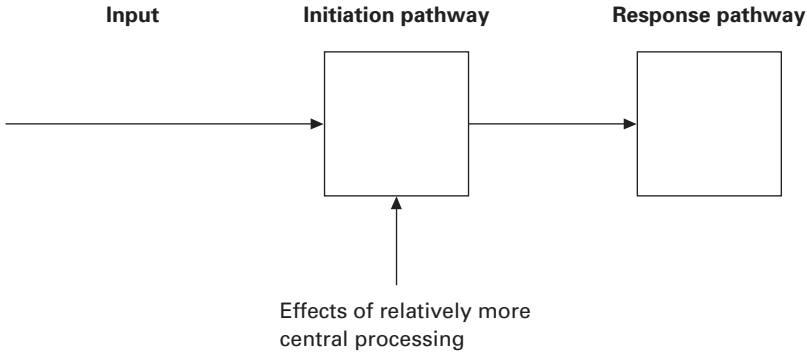
Strong non-instrumental content dependence Effects on content are brought about by changes to the response pathway without either direct changes to input or indirect changes via modification of the calibration files.

Obviously, instrumental content dependence alone means that horizontal modularity lacks empirical support. Finding strong non-instrumental content dependence would provide empirical support for abbreviated horizontal modularity. What about weak non-instrumental content dependence? The significance of this depends upon one's view of recalibration and input. If, on a vertically modular view, recalibration must happen through the same input channels as normal processing, then weak non-instrumental content dependence is consistent only with horizontal modularity. But if a vertically modular view recognizes recalibration via non-input pathways, then weak non-instrumental content dependence is consistent with both vertical and horizontal modularity. Given a tendency to assume vertical modularity, or at least the general view of the mind in which vertical modules have their natural home, discovering a state of affairs that, empirically, calls for abstaining from judging between

1. Instrumental emotional content dependence



2. Weak non-instrumental emotional content dependence



3. Strong non-instrumental emotional content dependence

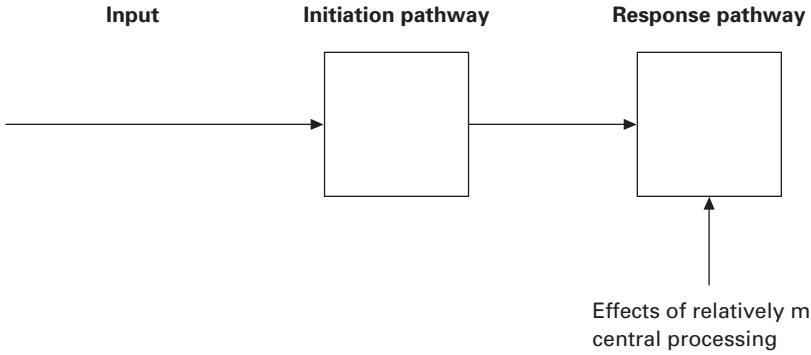


Figure 4.5

vertical and horizontal modularity would still be interesting, even if inconclusive.

Can the sorts of studies examined in this subsection shed much light on the likelihood of finding either weak or strong non-instrumental content dependence? More research seems to be needed on just what the input pathways to the amygdala are. As was the case for assessing the role of inhibition, studies of the development of our emotional capacities and of emotional pathologies are particularly promising sources of information on this topic. Once we are clearer about the input pathways, this information can be compared with the routes by which emotional regulation occurs. Ochsner et al. (2002, 1223–1224) speculate about three routes by which the LPFC might modulate the activity of the amygdala:

Directly. This is unlikely, since there seem to be few direct connections between these regions.

Via the MOFC. This is also unlikely, since the correlation between LPFC activity and MOFC activity was not as significant as that between LPFC activity and activity in the amygdala.

Via the occipital and parietal regions.

Overall, Ochsner et al. think more research is needed on this issue. If it turns out that the routes by which the LPFC affects the amygdala are much the same as the general input pathways to the amygdala, then we would have empirical support for either vertical modularity or, perhaps, weak horizontal modularity. But if the routes by which the LPFC affects activity in the amygdala differ from the normal input routes, then we would have empirical support for weak or even strong horizontal modularity of the relevant emotion.

Reflections

What does all of this mean for the Tempting View? Overall, this discussion has been aimed at the general view of the mind used by Prinz in his discussion of emotional perception. Although it makes sense to see the reactive attitudes in terms of emotional perception, considerations of the empirical assessment of vertical versus horizontal modularity oblige us at least to withhold judgment about this general view. The *a posteriori* grounds for confidence in either sort of modularity as a general structure of the

mind seem not to have been provided yet. With regard to the Tempting View, this means that we should not commit ourselves to its sufficiency claim: it has not been demonstrated that the mechanisms of emotional perception in general, and of the reactive attitudes in particular, are constitutively distinct from their various modes of expression. The most concrete findings surveyed in this territory—about mirror neurons and the mechanisms of empathic emotional recognition—point toward horizontal modularity and away from vertical modularity. The obvious pluralistic middle ground makes for a reasonable hypothesis: some instances of feelings alone suffice for attributions of responsibility but others do not since these others are not constitutively distinct from their expression.

It is worthwhile to stand back from these details and to think about the reactive attitudes themselves. In some ways considerations of the constitutive interdependence characteristic of horizontal modules are particularly germane to discussions of the reactive attitudes. Modeling them after classical perceptual modules is reasonable in view of the importance of the processing of input to the sorts of reactions to which Strawson draws our attention, but it also risks obscuring an important aspect of these attitudes. Arguably the reactive attitudes are just as much a means of producing behavior as they are for processing input. After all, Strawson's classic discussion brings them up to account for attributions of responsibility, which is something agents do, primarily toward other agents. These attitudes straddle the conceptual distinction between input and output which is theorized as a significant psychological division in the classical sandwich view of the mind. If vindicated by future empirical results, the rejection of the classical sandwich view ought to provide a foundation for modeling of the reactive attitudes that gives both its input and output aspects equal footing.

4.8 The Necessity Claim of the Tempting View: Reactive, Enactive, and Symbolic Cognition

The next step in this reconsideration of the Tempting View of the reactive attitudes is aimed at its necessity claim, i.e., that having feelings of a certain kind is necessary for the attribution of moral responsibility. This gets us to more explicitly externalist themes than the preceding discussion of emotional perception. The starting point is provided by Rob Wilson's distinc-

tion between reactive, enactive, and symbolic forms of cognition (2000, 38–40; 2004, 184–187).[9] Reactive representational systems are very closely linked to environmental signals, and hence the behavioral effects realized by such systems are "effectively under the control of the stimuli in [their] environment" (Wilson 2004, 185). Wilson's example of human reactive representational systems and behavior is reflexes (186). Enactive cognition gives the organism more control over the range of behavioral responses to input. Unlike reflexes, enactive cognition does not automatically produce a response of a particular kind. Wilson offers bodily skills, such as those required to ride a bike, as realized by enactive cognition. Finally, the higher forms of cognition are symbolic. This is the most familiar kind of cognition; examples include "thought, inference, reasoning, planning, wishful thinking, and reflection" (186). Symbolic cognition is freed from its bodily origins by the symbolic resources it uses, such as language. Wilson argues that as we move from reactive to enactive to symbolic cognition, the systems involved are realized by wider and wider systems. Whereas he contends that reactive cognition is realized by the brain, Wilson thinks that enactive cognition is realized by a system constituted by the brain plus the body, and that symbolic cognitive systems are constituted by the brain plus worldly cognitive resources beyond the physical boundaries of individual organisms. To make this case, Wilson reviews research programs studying memory (2000, 40–43; 2004, 189–198), cognitive development (2004, 198–206), the role of culture in cognition (2000, 46–50), and, most significantly for our purposes, mind reading (2000, 44–46; 2004, 206–210).

Except for mind reading, I will not review the details of Wilson's discussion, and I shall postpone the look at mind reading until later in the chapter. For now I shall use Wilson's tripartite division to categorize ways in which we express attributions of responsibility. Some expressions fit very well into the general category of reflexes. When we feel immediate and angry resentment at somebody's lack of consideration of our safety, others might well recognize this merely from the color and arrangement of our facial features. Shame is closely associated, at least in English, with hanging one's head, and I take this to be more like a reflex, like the jerking of one's knee in response to a physician's tap, than like the skills involved in riding a bike.[10] Many of the examples in the passage from Christian Smith allow of plausible and familiar interpretation as reflex reactions.

However, other expressions are less like reflexes and fit better into the category of enactive cognition. Consider this example: A motorist cuts me off as I cross the road in a marked crosswalk. In response, I give him the finger. This gesture is produced quickly but clearly is under my control in a way that flushing with anger (which might well also occur in this situation) is not. I take this gesture to be, in part at least, one way in which we attribute blame to each other. Given that the gesture is, partly, a miming of an action that I am figuratively wishing upon the inconsiderate and dangerous driver, it involves the body in a way that more purely symbolic expressions—e.g., utterances—do not. Such gestures are examples of expressions of attributions of responsibility produced by enactive rather than reactive or symbolic cognition.

The hallmarks of enactive cognition are its differences from reactive and symbolic cognition. It is under an agent's control to a greater degree than reflexes, and it is tied to bodily movements to a greater degree than reflection, reasoning, or wishful thinking. In short, enactive cognitive systems are constituted by bodily movements decoupled from environmental stimuli. Besides gestures, facial expressions can also fall into this category. The production of a smile or a frown usually is a reflex, but it need not be one. Consider a parent who thinks her child deserves punishment for conduct of some kind, but who also finds the situation funny. Such a parent might fix her face into a frown in order to deliver the message of punishment, while "genuinely" smiling to herself in private. In this case the parent's facial expression is a bodily movement under her control rather than that of the environment, and hence fits into the category of enactive cognition. I take it that phenomena of this sort are familiar and common. Steven Levitt and Stephen Dubner remind us of its political mobilization in *Freakonomics*: Stetson Kennedy of the Anti-Defamation League gave the name "Frown Power" to the ADL's recommendation that people distinctly frown when they encounter bigoted speech (2005, 58). I take it that this program recognizes that people need not be upset about something to use the facial expressions characteristic of such feelings.

The possibility of such decoupling is even more extensive with symbolic cognition. For present purposes, consider the attribution of moral responsibility with language. Language provides an environmental cognitive resource beyond the physical bounds of the agent. Certainly, feelings of indignation or disapproval can be expressed in language, either spoken or

written. But language can also be used to attribute responsibility in the absence of these feelings. Consider the amused parent and the Frown Power advocate. Besides fixing a frown on her face, the amused mother can attribute responsibility with a stern lecture. Frowning, whether expressive of present feelings or of more official judgments, is of use only in the actual presence of the bigoted speaker. Other measures have to be used when one encounters written bigotry or audio-visual recordings of bigots. This goes whether one actually feels a distinct reactive attitude or not. In such cases, letters, blogs, emails, telephone calls, and other modes of public speech can serve to attribute responsibility, and all of these can be decoupled from the feelings on which Strawson focuses.

All of this challenges the necessity thesis of the Tempting View. In view of the possibility of decoupling of expression from feeling in both enactive and symbolic attributions of responsibility, it seems simply false to think that certain feelings are necessary for such attribution. However, it is possible that the capacity for such feelings is necessary for the development of the capacities involved in producing the decoupled expressions. This notion preserves the spirit of the necessity thesis of the Tempting View, although it jettisons it in letter.[11]

4.9 Experimental Philosophy and Intuitions about Responsibility

Let's think about symbolically encoded attributions of responsibility a bit more. The reflections marshaled above present this phenomenon as psychologically heterogeneous. On one hand, we can use symbolic means of attributing responsibility when we actually have the feelings characteristic of the reactive attitudes. On the other hand, these feelings do not appear to be necessary for attributions of responsibility: we can use symbolic means of blaming and praising when we both recognize that indignation and gratitude are appropriate and lack the feelings themselves. Besides Strawson's work and the lay reflections gathered so far, is there any empirical support for this heterogeneous picture?

Symbolically encoded attributions of responsibility have recently been studied by experimental philosophers. This work supports the heterogeneous view presented here, and so gives us more principled reason to reject the necessity claim of the Tempting View than we have thus far had. As we saw in chapter 3, experimental philosophy begins with traditional

philosophical appeals to intuitions. Instead of acquiescing in a philosopher's own intuitions about a given topic, experimental philosophers design questions to assess the pre-theoretical intuitions of ordinary people about the topic in question. This has been done with a variety of topics, including moral responsibility.[12]

Studies of moral responsibility, and the closely related topic of free will and determinism, are found in Nichols 2004b, Nahmias et al. 2005, Nichols 2006, and Nichols and Knobe 2007. Both a review and a discussion of results are presented in Knobe and Doris 2010. The topic in the studies by Nahmias et al. is the compatibility of moral responsibility and determinism. Nahmias and colleagues presented subjects with scenarios involving a supercomputer and complete information about the laws of nature. Here is one scenario in most of its original detail:

> Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25, 2150 AD, 20 years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00 pm on January 26, 2195. As always, the supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 pm on January 26, 2195.
>
> Do you think that, when Jeremy robs the bank, he acts of his own free will? (Nahmias et al. 2005, 566)[13]

When asked whether someone could rob a bank, go jogging, or save someone from a fire of their own free will in such a world, subjects gave compatibilist answers: 76 percent said Jeremy robbed a bank freely, 68 percent said he saved the child freely, and 79 percent said he jogs of his own free will (Nahmias et al. 2005, 566–567). With regard to moral responsibility, the numbers are equally important: 83 percent said Jeremy was responsible for robbing the bank, while 88 percent judged that he was responsible in the positive case of saving the child (ibid., 568). However, Nichols (2006) has found evidence that, although people have deterministic intuitions, they also have indeterministic intuitions, which suggests that under different conditions these intuitions could be mobilized to give incompatibilist answers about moral responsibility and determinism. Against this background, Nichols and Knobe (2007) ran tests that tested

the importance of concrete versus abstract descriptions of cases. Concretely described cases provided descriptive detail particular scenarios, but abstractly described cases did not. Here are their cases in their original detail (source: Nichols and Knobe 2007, 669–670)[14]:

Background: Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to have French Fries.

Concrete Scenario: In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family.

Is Bill fully morally responsible for killing his wife and children?

Abstract Scenario: In Universe A, is it possible for a person to be fully morally responsible for their actions?

For the concrete cases, 72 percent of subjects judged that people could be morally responsible in a deterministic setting. But in the abstract case, 86 percent gave incompatibilist responses. Nichols and Knobe hypothesize that affective capacities are activated in the concrete case to a higher degree than in the abstract case. The exact details of the psychological capacities involved have still to be worked out, but it should be clear that should they receive further support, these findings would support the view of the psychology of the attribution of moral responsibility presented in this chapter. Some attributions of moral responsibility involve our affective capacities to a high degree, whereas others seem not to.

It is worth noting in passing that the findings of experimental philosophy regarding moral responsibility pertain to symbolically encoded attributions only. These studies present subjects with linguistically encoded descriptions of cases and solicit linguistically encoded responses. The psychological mechanisms of reactive and enactive attributions of responsibility are not directly tested by such means, and hence no direct conclusions about these other ways of attributing responsibility can be drawn from the

extant work done by experimental philosophers. In view of their methods, we should not expect any directly relevant work to be forthcoming either.

4.10 Mind Reading and the Reactive Attitudes

Let's stand back from the discussion so far to gather our bearings. We began with a reconstruction of Strawson on the reactive attitudes and moral responsibility, which I called the Tempting View: taxonomically wide but locationally narrow feelings of a certain kind are both necessary and sufficient for the attribution of moral responsibility. Against this, we have seen, on one hand, empirical challenges to the background view of the mind in which this view finds its natural home; on the other hand, we have seen a broadening of our catalog of ways of attributing responsibility beyond the reactive attitudes themselves. Hopefully the fruits of this investigation are a more realistic view of the psychology and practice of the attribution of responsibility than that provided by the Tempting View. The cost, however, is psychological simplicity: whereas the Tempting View explained the attribution of responsibility in terms of a single, homogeneous class of attitudes, the picture that replaces it retains this class and adds other psychological capacities of varying degrees of locational width. Nothing in particular seems to unite these capacities. Is it really the case that the psychology of responsibility attribution is radically heterogeneous? Or is there something in common to all forms of responsibility attribution that has so far been neglected?

I think that there is a unifying psychological thread to the attribution of responsibility, and that it is present, although underdeveloped, in the original Strawsonian position. When developing the Tempting View, I said that the reactive attitudes were triggered by and directed toward certain sorts of thoughts. This implies that these attitudes require whatever psychological capacities are used to understand the minds of others. I think that this is broadly correct, but that, again, it requires deep modification.

To frame the crucial issues, let's attend to some distinctions using Strawson's account as our guide. For Strawson, the reactive attitudes are deployed in response to the intentions and attitudes of others. There are two ways in which we can respond to the mental states of others. First, they can be the *trigger* or the *cause* of the reactively attitudinal response. Second, they can be the *grounds* or the *warrant* for the reactively attitudinal response. I

presume that they can fill these roles simultaneously. Consider a scenario in which a driver laughingly cuts off a pedestrian at a marked crosswalk. Here we have relatively clear evidence of callousness toward someone's rights and well-being. Suppose that the pedestrian responds to such an obvious slight and threat by giving the motorist the finger. It is reasonable to see the motorist's activity as both the causal trigger of the pedestrian's response, and as at least part of the warrant for the response. That is, if asked why he made a vulgar gesture, the pedestrian could reasonably cite the attitudes of the driver as evinced by the driver's behavior. In this case, not only do the driver's attitudes play the role of cause and warrant of the reactive attitudes; they are also central to the *target* of these attitudes: the pedestrian's gesture is directed toward the driver's callousness in particular, and perhaps toward the driver's character in general.

On the Strawsonian story, mental states are the only things offered as filling the roles of trigger, warrant, and target of the reactive attitudes.[15] However, it is reasonable to think that real attributions of responsibility are more complex than this. Things other than the thoughts of others can play the roles of trigger, warrant, and target of attributions of responsibility. To see this, consider a variation of the pedestrian-driver case. Instead of laughingly cutting off the pedestrian, suppose that the driver is preoccupied with other issues and cuts off the pedestrian by accident. In such a case, the trigger of the pedestrian's response is not a particular thought that the driver has: the driver's thoughts have nothing to do with the pedestrian, nor even with driving, and hence are not directed at the pedestrian. Instead, the trigger is the driver's action. The warrant for the pedestrian's response might well have something to do with the driver's psychology, but it would be a mistake to charge the driver with callousness. There is no direct evidence of such an outlook. Instead, the driver is guilty of *lacking* appropriate mental states—that is, of failing to attend to the task at hand and the dangers it poses to others who have a right to use the same roadways. The target of the response is, in all likelihood, the same between the two cases: the driver's character. Similarly, some of the examples in the passage from Christian Smith are plausibly seen as triggered by or targeted at actions. The wife's annoyance is triggered by her husband's action of watching television, for example.

In principle, all three roles (trigger, warrant, target) can be played by things other than psychological states. To make this case, and to add a sort

of real-world weight to the consideration in this subsection, let's turn to an actual case of the attribution of responsibility. In Canada, July 1 is Canada Day, our primary patriotic holiday. People get a day off from work, and there are celebrations of various sorts across the country. The most prominent celebrations take place in the national capital, Ottawa, where I happen to live. Fireworks and concerts take place at and around the Parliament Buildings in downtown Ottawa. Also nearby is the national war memorial. As one would expect, there is celebratory drinking of alcohol at these summer events. On July 1, 2006, these factors conspired to produce a minor national incident. Some young men celebrating Canada Day downtown were drinking and watching the fireworks. They needed to urinate, but they did not want to miss any of the show. Once the fireworks were done, the men's need to urinate was becoming urgent. Drunk and not realizing the significance of what they were doing, they relieved themselves on the side of the national war memorial. As it happens, some veterans had suspected for some time that the memorial was not protected as well as it should be, and had taken to keeping an eye on it as a sort of voluntary surveillance service. They recorded the men's activity, then told the media and the police. For a few days, there was notable public outrage across the country. The men were identified and charged with mischief. They issued public apologies in which they said they hadn't known what they were doing—they were drunk at the time, and they were ignorant of the identity and significance of the item on which they relieved themselves. They saw it as either a wall or a rock, not as something dedicated to officially recognizing the lives lost by Canadian soldiers in international wars. They disavowed any intention to insult.

What should we make of all this? First, this offers a very clear case of attribution of responsibility via the reactive attitudes. Second, we should take the men's public apologies at face value: it is very plausible to think both that they had no intention to insult and that they were ignorant of the significance of the monument. I think this means that we should see their acts as providing the warrant for the reactive attitudes. They are also their trigger. The tricky part is determining the target of the reactive attitudes. It is less appealing to see the young men's character as the target in this case than it was in the driver-pedestrian cases. For one thing, before urinating, they were acting in a manner that is broadly accepted, even encouraged, in Canada. Mild public drunkenness is widely accepted, as is

taking part in Canada Day celebrations with alcohol. All this makes it unlikely that the reactive attitudes should be interpreted as personally directed in this case. This case provides a uniquely clear example of one version of the generalized attitudes that Strawson presents as the particularly moralized exemplification of the reactive attitudes: the public response was directed at the men's actions and at the general fact that something such as this could happen at all.

Here is a different way of putting the spirit of this discussion: It is reasonable, I think, to see the moral issue in the men's actions in terms of disrespect. In some forms, respect and disrespect are a matter of having and exhibiting certain sorts of thoughts about people and objects. However, I think that the war memorial example presents a case of *thoroughly objective* disrespect, and a response to it. There was something morally problematic about the men's actions *regardless* of their thoughts about them. A sincere and apologetic report of their attitudes toward the monument should suffice to allay worries about mentalistic disrespect. However, they do not address objective disrespect. In this case, the disrespect is brought about by the performance of a conventional symbol of disdain; this is the case even after the men sincerely apologize. The public response is reasonably interpreted as partly about this conventionalized symbol. If this is correct, then the target of the reactive attitudes in this case is neither the men's character nor anything to do with their psychology, but rather the public and thoroughly objective disrespect that their actions instantiated.[16]

Let's say that an instance of the reactive attitudes that is triggered, warranted, *or* targeted toward the thoughts of somebody is a *mind-reading-directed* attitude, or MR-directed for short. Attitudes that are not triggered, warranted, or targeted toward someone's thoughts are *non-mind-reading-directed* attitudes. Most reactive attitudes are MR-directed, but those found in the Canada Day war memorial case are non-MR-directed. This seems to present us with continuing heterogeneity. However, although Strawson is incorrect to present moral responsibility as always MR-directed, he is not wrong about the importance of mind-reading capacities to moral responsibility. As it happens, even non-MR-directed reactive attitudes require mind-reading capacities. Consider: non-MR-directed reactive attitudes are directed at people's actions.[17] Identifying an event as an action rather than as a non-action seems to require the ability to identify someone's

intentions. Identifying someone's intentions is a particular sort of mind reading, so even the non-MR-directed reactive attitudes draw on our mind-reading capacities.

Although the present topic is the mature capacity to attribute responsibility, here are some very brief developmental considerations in support of the idea that action-identification requires some sort of mind-reading capacity. From very early ages, children imitate actions and expressions. For at least some actions, being able to imitate them requires being able to identify the thoughts that were guiding them. In an important and well-known study by Andrew Meltzoff (1995), infants watched adults try *and fail* to perform actions such as taking apart a toy dumbbell. Infants allowed to play with the dumbbell afterward were just as likely to take it apart as infants who watched an adult successfully take it apart. The interpretation here is that the infants who watched the adult fail to take the dumbbell apart imitated the adult's aim, not the adult's movements, and this requires understanding of the adult's mind. Such mind reading also seems to be involved in learning the meanings of words for objects and actions. For instance, children can learn the names of objects that are selected by adults in a series of objects. When adults search for an object with a particular name, in the process discarding some items before happily settling on one, children give the name in question to the happily chosen item, not the discarded ones. Doing so requires being able to understand that the adult is searching, which in turn requires being able to identify the adult's goals, as well as his or her feelings about the objects.[18]

Building on such considerations, it is reasonable to think that understanding people's actions requires being able to do a fair amount of mind reading. This entails that even non-MR-directed reactive attitudes require mind-reading capacities. This gives us the common thread to the reactive attitudes. As diverse as they are in many ways (see figure 4.6), they all require the ability to understand the thoughts of others. This provides us with an interesting way to approach the issue of the environmental dependence of the reactive attitudes. Let's put aside the externalist aspects of enactive and especially symbolic cognition and focus on mind reading. We have already cast our mind-reading capacities as taxonomically wide; is there any sense in which they are locationally wide?

In his discussions of externalism and mind reading, Wilson (2000, 44–46; 2004, 206–210) draws attention to two aspects of this topic. One is

	MR-directed	Non-MR-directed
Reactive Canada Day	Blushing in anger at someone's contempt for you.	Blushing in anger at the war memorial incident.
Enactive	Waving in gratitude for a driver's courtesy.	Brandishing one's fist when discussing the Canada Day war memorial incident.*
Symbolic	Lecturing someone for hateful remarks.	Writing a letter about the objective respect extended to veterans by programs to protect the Canadian national war memorial.

*This is the segment of the chart that I had the most difficulty filling; perhaps enactive attributions of moral responsibility are typically MR-directed.

Figure 4.6
Examples of the heterogeneity of the psychology of responsibility attribution.

that there is a difference between a rudimentary understanding of others as having such representational states as beliefs and desires and a richer understanding of others as having feelings, sensations, moods, and so on. Wilson thinks that the rudimentary mind-reading capacity may well be locationally narrow, but that the richer capacity is wide. For present purposes, although it is likely that the richer, more red-blooded capacity is regularly used in the reactive attitudes, it seems prudent not to suppose too much. Perhaps the identification of actions requires only the narrow belief-desire psychology, in which case this aspect of the reactive attitudes should be seen as locationally narrow. The second aspect is that mind reading involves not only understanding others but also interacting with them. In interpersonal interaction, the issue of the location of the locus of control of one's actions arises. In some cases, one clearly retains control. In cooperative endeavors that involve open and honest communication about ends and the means by which to attain them, rational individuals

maintain their practical autonomy. However, Wilson emphasizes the difference between such cases and those of manipulation and deception. In an important sense, when one's conduct is a product of manipulation or deception, then one has lost some degree of control over it. Someone else is the source of one's actions to an important degree. In such cases, the action-production system, utilizing mind-reading capacities, responsible for one's conduct is distributed between individuals, and is hence locationally wide.

How do the reactive attitudes and attributions of responsibility fit in here? In keeping with the flavor of this book, I think the answer is "heterogeneously." It is reasonable to think that we deploy responsibility in order to shape the behavior of others; I take it that this is not really in question. What is in question is in what ways this is done. Some attributions of responsibility are exactly like the open and honest discussion that I used to exemplify the retaining of control over one's action. Such attributions—perhaps especially when coolly, symbolically encoded—preserve autonomy. However, we should not underestimate the possibility of manipulation and deception that exploits the reactive attitudes. Emotional manipulation is familiar and effective; moralized emotional manipulation is a particular variety of this broader category. In such cases, agents lose the locus of control of their own actions, and hence we should see the action-production systems as locationally wide.[19]

I think there is room for middle ground here—that is, for cases in which one neither solely retains nor loses the locus of control of one's actions, but instead shares it. This is a reasonable interpretation of cases of attributions of responsibility that are neither solely for directing another's behavior nor for manipulating it, but instead for coordinating behavior between oneself and others. When aims are openly shared, and perhaps especially when these aims are not achievable by individuals acting alone, and when the reactive attitudes are deployed to coordinate behavior, I think we should see the locus of the overall behavior as distributed among the individuals who bring it about. I predict that this phenomenon could be found in the behavior of organizations—corporations, charities, and military forces come to mind. In all these cases, the ends of the group are achieved by individuals working together, they are not achievable without such coordinated activity, and emotional resources are used to bring about the coordination necessary to attain the ends in question. However, this analy-

sis of the role of attitudes in the coordination of behavior requires empirical investigation that has yet to be done.

These reflections pertain to the output aspect of the reactive attitudes, not to the input aspect. We should acknowledge the possibility that these are deeply interwoven. Recall the discussion of the structure of empathy and mirror neurons in chapter 2. This suggests that the same aspects of the brain play roles in the recognition of fear in other people and in the production of fear for the person experiencing it. If this pattern is found for the emotions characteristic of the reactive attitudes, then we have the beginnings of an indirect case for the locational width of their input processing. Here is the schema of the argument in very brief form: Since the action-production processing of the reactive attitudes is at least partially locationally wide,[20] and since the same structures realize the experience of the reactive attitudes and the recognition of the reactive attitudes in others, in order to play a role in the wide action-production system characteristic of such attributions of responsibility one also has to be able to experience and deploy the reactive attitudes toward others. This means that both are at least partially locationally wide. At present this is no more than a suggestion, but it is a tantalizing one for future investigation.

Finally, what about *being* morally responsible? In typical post-Strawson discussions, to be morally responsible is to be the kind of thing to which it is appropriate to direct the reactive attitudes. The present discussion suggests one thing that this might amount to: getting into the widely realized action-production system(s) that utilize(s) the reactive attitudes. If one is not included in these systems, then it seems that one is closed off from one way that deployment of the reactive attitudes can serve to influence one's behavior. If this is correct, it marks a change from the position on being morally responsible that I have previously defended. In my 2005 paper, I argued that that the psychology of moral responsibility is plausibly seen as taxonomically wide but locationally narrow. If the present case is correct, however, then an important aspect of the psychology of morally responsible agents is locationally wide too.

This sort of view provides tools for fleshing out some remarks made by Michael Gazzaniga. In a discussion of neuroscience and free will, Gazzaniga briefly considers moral responsibility. He sees it as neither threatened by nor apt to be illuminated by neuroscience. His reasoning is that moral responsibility is not a property of the brain. It is instead a property of entire

persons. Moreover, rather than being a phenomenon apt for inclusion in some branch of the biosciences, Gazzaniga calls moral responsibility a "social construct" (2005, 101–102). This sort of terminology is often used to downgrade the ontological status of something: it's not really real, it's a social construct, meaning an arbitrary result of social life that can be jettisoned from any serious catalog of real phenomena. Gazzaniga may mean this, but since he takes moral responsibility seriously this interpretation of his remarks is not particularly attractive. Nor is it the only one at hand, as we can now see. If our social lives provide the realization base for some psychological systems, and if some of these systems are at least partly constitutive of moral responsibility, then responsibility can be social and just as real as other phenomena realized by psychological systems that happen not to be socially realized. If this is the case, we can agree with Gazzaniga that responsibility "does not exist in the neuronal structures of the brain" (2005, 102) without even risking downgrading its ontological status.

Concluding Reflections

I began this chapter by reconstructing Strawson's work on the attribution of moral responsibility in terms of an individualistic position that I called the Tempting View. On the basis of a variety of kinds of consideration at both the performance level and the psychological level, I argued that we should reject the Tempting View. Instead of seeing responsibility as solely and necessarily attributed via certain sorts of internal experiences, we should construe this as a psychologically heterogeneous phenomenon.

Two things are worth emphasizing about this psychological heterogeneity. First, the additions to the psychology of the Tempting View are, to varying degrees, locationally wide. Not only is moral responsibility attributed using the resources of symbol systems, such as language, located beyond the physical boundaries of agents; it also affords a multi-faceted way of controlling and coordinating the behavior of others and of oneself. The action-production aspect of the attribution of responsibility is realized in a locationally wide system involving other agents. Second, despite its overall heterogeneity, all attributions of responsibility rely on our abilities to understand the minds of others. Though some attributions are overtly directed toward or by the mental states of others, others are not. However,

even those attributions of responsibility that are not directed by the mental states of others require recognition of their actions, and this requires understanding of the mental states of others.

To put it summarily, according to the Moral Systems Hypothesis, the heterogeneous psychology of the attribution of moral responsibility is partially realized by wide cognitive systems. Recall the schema from chapter 1:

_____ systems must be causally and functionally integrated chains of _____ resources, and these, individually and collectively, must play a replicable causal role in _____

This chapter has focused primarily on external resources; the resources in question for attributions of responsibility are those provided by such external symbol systems as language and, most important of all, by the minds of other people. I hope to have delivered the possibility of this aspect of the WMSH in this chapter, and maybe even its initial plausibility.

5 The Production of Action

Recent years have seen the development of the implications of the "person-situation" debate in psychology for philosophical discussions of virtue, most notably by Gilbert Harman (1999, 2000) and John Doris (1998, 2002). The reception of this work has been lukewarm at best. I, for one, have been convinced, so I find this a bit puzzling. Presumably one reason for this state of affairs is that philosophers have not been convinced by the case that has been presented. Although I will briefly review this case, I have little to add to it. A deeper reason why the "situationist" case has not been well received is that both those making the case and those resisting it have underestimated the scope of the implications of this work. This will be the principal theme of the present chapter. To put it as straightforwardly as possible, I will return philosophical discussion of this debate to its original topic: the production of action. Since this is a book about *moral* psychology, I will begin with the familiar philosophical debate about the psychology of virtue. However, my primary topic is a much wider one. The aim of this chapter is to develop the Wide Moral Systems Hypothesis by demonstrating the possibility, the plausibility, and the moral-psychological importance of an externalist position on the production of action.

Here are two vignettes to introduce the issues.

First vignette: Michael is watching television. He rises from his seat and leaves the room. Kim asks "Why did John go to the kitchen?" Ramona replies "He wants a beer and knows that there is some St. Ambroise Oatmeal Stout in the fridge." This reply answers Kim's question by citing two sorts of psychological state: a *desire* for a beer and a *belief* that there is a beer in the fridge.

For decades many philosophers have thought that an important way to explain actions—a way that may be the very core of any adequate

explanation of action—is in terms of the combination of states (e.g. beliefs) that represent the world and other states (e.g. desires) that realize a person's values, goals, and, most generally, wants. Such explanations are common and seem to work. It is reasonable to think that when such explanations accurately explain behavior, it is because they have accurately captured something about the processes that produced the action in question. This means that how we understand explanations of action is a guide to the ways in which action is produced. The adequacy of belief-desire explanations of actions will not be radically questioned here. The present question is how to understand the states cited in an explanation of action. Must they be located within the physical bounds of an agent's body, or can they extend beyond these bounds to include parts of the wider world. Mostly implicitly, most philosophers and psychologists assume an individualistic view of the states offered in explanations of action, and correspondingly of the states thought to produce action. In contrast, I will defend a wide view of the psychology of action-production. I will discuss Donald Davidson's influential view of the explanation of actions in section 5.8. In subsequent sections I will offer an externalistic account of action production using Davidson's schema.

Second vignette: Kim is watching television. She rises and leaves her seat. Michael asks "Why did Kim go to the kitchen?" Ramona responds "She wants to use the phone in there. She just saw an advertisement calling for emergency aid to China. Kim's really nice, you know? I'd trust her with anything! She took care of my cat last summer. Now she's giving money to Oxfam."

This explanation of action is much like the first. It consists of a belief that aid is needed in China and a desire to give money to help are cited to explain why Kim leaves the room. But in this story there is also an additional component: a trait of character—Kim's benevolence—that by many standards would count as a virtue. I take it that there is nothing odd about finding such a trait in an explanation of behavior. The corresponding idea is that such traits can function in the production of behavior. I also assume that there is nothing peculiar about Ramona's generalization of Kim's niceness on the basis of two instances. There is reason to think that a common view of individual psychology describes people in terms of character traits that are supposed to function across a wide variety of different situations. Character traits such as benevolence, and the virtues and

vices in general, are widely thought to affect how people see things and, crucially for present purposes, how they act very generally. If you know someone in your workplace who is nice, presumably this gives you information that is useful in other contexts—a mall, a tennis court, a theater, and so on. This assumption has been the focal point of the so-called person-situation debate. I side with the situationists in this debate: I think the assumption is false. In lieu of traits that apply anywhere, situationists offer a much more context-sensitive account of our psychology. Someone who acts nicely in a workplace need not have the psychological capacities that deliver nice actions elsewhere. I take such context sensitivity to be a clue that our action-production capacities are widely realized.

Before getting into the psychology of the production of action, I shall look at the person-situation debate and the psychology of virtue.

5.1 The Person-Situation Debate

In this section, I provide a map of the person-situation debate. The positions are thinly described to convey the sense of poles of a debate characterized by much mapping of the territory in between. Given the simplifications, proponents of both sides may well find things to object to in this characterization. Still, my hope is that this introduction is more useful than distorting.

The person-situation debate was sparked by Walter Mischel's review of the literature on personality and action production in *Personality and Assessment* (1968). After Mischel's work, some social psychologists (sometimes called "situationists") and some personality psychologists carried out a vigorous and interesting discussion of the relative contributions of personality structures and the environment to the production of behavior. Very roughly, personality psychologists argue that variation in behavior between individuals is due to variation in certain sorts of psychological traits possessed by those individuals. Exactly what sorts of psychological traits is one part of the debate. For example, if you are interested in explaining why Meghan succeeds at school but has few friends, and why Josef fails academically but has many successful personal relationships, the broad approach offered by personality psychology directs you to find out about certain character traits that putatively produce behavior. As a specific example, the Five-Factor Model of personality—a model associated with

work done by Lewis Goldberg (1993) and by Robert McCrae and Paul Costa (1996)—presents personality as composed of five traits, sometimes given the names Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. The general idea is that differences in these five factors account for differences in personality and hence for differences in behavior. Five-Factor theorists would attempt to explain the differences between Meghan and Josef in terms of these traits. Correlatively, if you were interested in predicting how Neville and Joni would perform in a particular institutional setting, personality psychology would advise you to find out about their personality traits, perhaps according to the Five-Factor model, perhaps by the standards of a different list of personality traits. In contrast, and again very roughly, situationist social psychology argues that variation in behavior is due much more to differences in situations than we are inclined to think. Particularly important for present purposes are studies about the production of behavior that is characterized in explicitly moral terms or in situations that are characterized in explicitly moral terms. *Studies in Deceit*, a 1928 book by Hugh Hartshorne and Mark May, stands at the beginning of the twentieth-century situationist tradition. Hartshorne and May performed a long-term study of deceit involving thousands of children in classroom settings. They used a variety of tests to assess their subjects for deception and honesty in various forms, such as cheating on tests or lying to teachers. Their finding that the correlation between different sorts of honest behavior or deceptive behavior was remarkably low led them to infer that the variation in behavior was better explained by variation in properties of the immediate context than by some sort of personality trait.

Subsequent studies provided evidence in support of this idea. Stanley Milgram's infamous studies on obedience (1963) are the best-known of these. Milgram conducted studies that putatively were about learning but actually were about obedience to authority. Subjects were given the role of teacher in these studies; confederates of the experimenters played the roles of learner and study administrator. The teacher's jobs were to ask questions and to administer electric shocks in response to incorrect answers. The shocks ascended in severity in 15-volt increments. Some levels of shock were clearly labeled with fairly dire warnings. When subjects hesitated in administering shocks, the administrator-confederate politely recited a list of instructions to continue. Milgram found that non-coercive features of

experimental situations led ordinary people to administer what they thought were lethal levels of electrical shocks to other ordinary people. More precisely, about two-thirds of subjects administered shocks all the way to the final level, and many of the other subjects administered shocks up to very high levels.

Other studies assessed helping behaviors rather than harming ones. Here is an example that Doris emphasizes as illustrative: Alice Isen and Paula Levin (1972) found a very high correlation between the performance of helping behavior and seemingly insignificant good fortune, such as finding a dime in the change slot of a pay phone. In their experiments, subjects (unsuspecting ones, not solicited ones) were people who went into a phone booth. Some found a coin in the change slot, others did not. When they left the booth, an experimental confederate posing as a passerby dropped a pile of papers, apparently accidentally, outside the phone booth. Of the 16 people who found coins in the phone, 14 helped and 2 did not. Of the 25 people who did not find coins, only one person helped with the dropped papers. Doris (2002, 34) reports that more than 1,000 studies have produced results like these about helping behavior alone.[1]

Overall, the situationist suggestion is that the variation in behavior exhibited by an individual should be accounted for in a way that gives a substantial role to variation in context. Just what sort of role, and just what sort of contribution is made by individual psychology, again figures in the debate. To explain the differences between Meghan and Josef, situationist social psychology would look at least in part to the contexts in which Meghan and Josef perform. For the purposes of predicting how Neville and Joni will behave within a given institutional setting, situationist social psychology *might* direct us to find out about the details of this context, but it would more likely advise us to find out about how people generally behave in such settings.

5.2 The Person-Situation Debate and the Renaissance of Virtue

The principal philosophical target of Doris and Harman is appeals to virtues and vices, which have gone from bit players to marquee stars in ethical theory over the last half-century. This renaissance of interest in the virtues can be traced back to Elizabeth Anscombe's 1958 paper "Modern Moral Philosophy." But when Anscombe advocated that philosophers stop

doing moral philosophy "until we have an adequate philosophy of psychology," we can be fairly sure that she did not foresee the naturalistic and interdisciplinary turn that occurred in the philosophy of mind in the late twentieth century. Moreover, we can be absolutely sure that when she recommended a revival of virtue-theoretic notions in moral philosophy, Anscombe had no idea that certain threads in the empirical study of personality and the production of behavior would call into question the very existence of such psychological traits as the virtues. Yet that is exactly what happened.

How does the person-situation debate connect to philosophical use of the ideas of virtues and vices? Let's begin with everyday versions of these notions. Typically included on lists of virtues are character traits such as honesty and courage. Typical vices are greed and cowardice. These are generally taken to be constituents of different types of personality. Moreover, virtues and vices are typically assumed to be centrally involved in the production of behavior. That someone is courageous is often taken as centrally explanatorily relevant of that person's behavior in various sorts of situations. If you want to predict how someone will behave, you might well try to learn whether he or she is courageous or cowardly. Overall, we have reason to think that folk appeals to virtues and vices correspond to the general account of persons and behavior offered by personality psychology. Situationist social psychology offers a different view of our behavior from that presented by personality psychology, so if situationism turns out to be true, then our everyday notions of virtue and vice will correlatively turn out to constitute an inaccurate view of the nature of persons. The same would go for the folk explanations of behavior in which these everyday notions of virtue and vice have their home.[2]

Even if one doubts that our everyday use of the notions of virtue and vice is personological, there is good reason to think that principled philosophical use of these notions generally is. A cursory flip through the pages of latter-day virtue theory provides plenty of evidence. N. J. H. Dent (1984, 9–10) thinks that all character traits are either virtues or vices. He offers kindness and conscientiousness as examples of virtues (17). Rosalind Hursthouse (1997, 220–221) also refers to virtues and vices as character traits. She adds the examples of benevolence and justice to the putative virtues we have already seen. Linda Zagzebski (1996, 137) offers the following definition of virtue: "a deep and enduring acquired excellence of a

person, involving a characteristic motivation to produce a certain desired end and reliable success in bringing about that end." That virtues are traits of persons responsible for behavior is explicit in this definition. Besides the examples already presented, Zagzebski counts as virtues fairness, self-improvement, and generosity (1996, 113).

The psychological challenge to both everyday and more principled appeals to virtue can be put succinctly by asking the following question: What grounds our confidence in the existence of virtues and vices? The most likely answer, it seems to me, will appeal to their role in explaining behavior. But this means that principled inquiry into the production of behavior, such as that offered by psychology in general and by personality and social psychology in particular, is directly relevant to the reasons we have for thinking that people have such behavior-producing traits as courage and kindness. If such inquiry into the production of behavior reveals that such traits have little or no role, then our confidence in the very existence of virtues and vices would turn out to have no principled foundation. These ideas *might* do for everyday purposes, but they would deserve no role in philosophical theories of the production and evaluation of human behavior. Empirical psychology would be virtue's demise.

5.3 A More Detailed Look at the Psychological Challenge to Virtue

Let's take a closer look at the notions of virtue and vice and at the challenges posed to them by the person-situation debate.

First, consider the following schema about virtue and vice.[3] What I shall call The Traditional View of such character traits as virtue and vice, found in both folk and theoretical contexts, makes two assumptions:

Regularity Character traits, which are dispositions to produce behaviors, function in regular, patterned ways.

Autonomy When they produce behaviors, character traits operate with substantial independence from contextual contingencies.

For example, according to the Regularity assumption, an honest person exhibits particular sorts of patterns in behavior: such a person she tells the truth, doesn't lie, sticks to agreements, and so on. A dishonest person exhibits patterns of a different sort. According to Autonomy, the traits of honesty and dishonesty are the most important determinants of these

agents' behavior when it is legitimately describable in trait-relevant terms. The implicit view is that situation-independent traits are the most significant determinants of behavior *tout court*. Honesty suffices to produce the patterned behavior characteristic of the honest person. Autonomy entails that the psychological determinants of behavior are not indexed to particular types of situations. People are honest or dishonest, not honest-at-work or dishonest-in-the-car. Besides the examples we have already seen, consider John McDowell's neo-Aristotelian view of the virtues. This view shows these assumptions at work with an explicit focus on normative evaluation. McDowell (1997, 143) claims that virtues produce right conduct only. Being productive of right conduct only is a pattern, so McDowell's view exemplifies Regularity. It also exhibits Autonomy, since McDowell's position implies that it doesn't matter what sort of context the virtues function within. Place an honest person in an office, in a home, or on a street, and his honesty will always produce right behavior, provided that other features of the agent's psychology do not interfere.

The challenge to virtue posed by situationist social psychology is, crucially, two-pronged. Each prong addresses one of these structural assumptions of The Traditional View of character traits, but in a way that constitutes a complex, mutually reinforcing, and very serious challenge to this view. The first prong of the challenge arises from studies such as those of Hartshorne and May. For present purposes, the most striking of their findings is number 23: that deceit and honesty are not unified traits (1928, book I, 411–412). Hartshorne and May found remarkably low correlation between different sorts of honest behavior or deceptive behavior, which led them to infer that the variation in behavior was better explained by variation in properties of the immediate context than by some sort of personality trait. Forty years after the Hartshorne-May study, Walter Mischel's 1968 review of the literature on various types of behavior found the same sort of pattern across the board. Overall, the lesson seems to be that behavior is much less consistent than one would suspect if it were produced by unified traits of persons. That is, the pioneering studies that revealed substantial behavioral inconsistency called into question the Regularity assumption of the traditional view.

A potential reply might be based on the possibility of taking a more nuanced position on the composition of personality traits. Instead of seeing such traits as honesty and deception as relatively simple and

unstructured, perhaps we should treat them as complexes constituted by many different action-producing mechanisms. Assuming that these mechanisms can function independent of each other, inconsistency in behavior can be made compatible with the traditional view. This is also a way of accounting for the possession of virtue and vice in degrees—people can be mostly honest when most of the mechanisms that constitute honesty function properly, but a little bit of deceptive behavior produced by other submechanisms would be consistent with possession of this trait. On this view, instead of following Hartshorne and May and Mischel in rejecting traits in order to explain inconsistent behavior, the explanation is provided by a finer-grained view of the constitution of these traits. The details of this view would be filled in with further empirical investigation.

It is at this point that the second prong of the psychological challenge to virtue becomes important. The suggested reply to the rejection of traits in order to explain cross-situational inconsistency in behavior exemplifies the Autonomy assumption: the mechanisms that purportedly realize virtues and vices operate independent of the vagaries of situations. If they did not, then they would not realize character traits as modeled by the traditional view. The studies by Milgram (1963) and Isen and Levin (1972) are only two of the many examinations of the role of context in producing behavior. These studies call into question Autonomy, and hence call into question this way of replying to the rejection of traits in order to explain cross-situational inconsistency of behavior. Recall that Milgram found that seemingly non-coercive features of experimental situations led ordinary people to administer what they thought were lethal levels of electrical shocks to other ordinary people. In contrast, Isen and Levin found a very high correlation between the performance of helping behavior and the seemingly insignificant good fortune of finding a dime in the change slot of a pay phone. In view of such findings, and those of Hartshorne and May, Mischel, and others, an attractive explanation is that the mechanisms that produce behavior are context specific, such that features of one's environment can be very significant determinants of one's behavior. On this view, personality structures are not the only source of patterns in behavior; features of an agent's environment are another source.

If the situationist studies are correct, then the two structural assumptions of The Traditional View of character traits are mistaken. Insofar as a psychological or philosophical theory requires an accurate view of the

production of human behavior, commitment to The Traditional View will compromise it. Since many philosophers are committed to The Traditional View by way of their theoretical deployment of traditional notions of virtue and vice, these philosophers must face the details of the challenge to virtue posed by situationist social psychology. They must also be careful about the conceptual resources found in the person-situation debate that may provide an answer to this challenge.

5.4 A Clash of Psychological Research Programs

A notable feature of the empirical tradition of situationist psychology is that it is overtly concerned with explicitly moral behavior[4] (such as helping and harming) and with situations presenting moral dilemmas (such as whether or not to cheat on a test, or whether or not to be obedient to a research protocol to the extent of harming or even killing another person). Another notable feature is how surprising these results are. Nobody predicted in advance that Milgram's subjects would behave as they did; even subjects themselves are poor predictors of how they would behave in such circumstances. Likewise, in general we do not tend to suspect high correlations between tiny good fortune, such as finding a coin, and helping someone. These two features, plus the demonstrated environmental sensitivity of the psychology of action-production, deliver a marked contrast between situationist psychology and the sorts of studies of moral judgment and moral reasoning I examined in chapters 2 and 3. In particular, they provide a perspective from which to evaluate the scope of the importance of the findings of these studies. I shall focus on the moral/conventional tradition, since it is the best-developed and the most famous body of work, but the point of this section can be extended to research programs with similar methods and assumptions. To put it the other way around, the tradition of studies of the moral/conventional distinction provides an implicit challenge to situationist psychology. I shall examine this clash of research programs in order to begin reflecting on the status of the findings of situationist psychology. I shall turn more explicitly to thought about virtue in the next section.

Recall that Nichols, Smetana, Turiel, and others focused primarily on moral judgment using studies of moral reasoning. Let's put aside doubts about the findings and assume that there is a robust psychological distinc-

tion between the processing of moral and conventional issues respectively. Our options for what to make of this fall along a range with the following poles: either the capacities disclosed by these studies are only parts of our overall moral psychology, with no particularly central role therein, or they are its very foundation, such that other psychological systems rely on these for their functioning. As should be evident, I favor the first option. Let's reflect on the second option. Adopting this option brings with it commitment to two theses:

Substantial Our central moral-psychological capacities are those that are used to give verbal responses to questions about hypothetical scenarios.

Methodological Our central moral-psychological capacities can be accurately studied via consciously accessible propositional knowledge that is deployed independent of the production of actions in real contexts and of real interactions with other people.

Let's begin with the substantial thesis. If this is correct, then either (A) studies of the production of action do not tell us anything about our central moral-psychological capacities or (B) the psychological capacities tapped in moral/conventional-distinction tests account for action-production, and so they should predict the results of empirical studies of the production of morally relevant behavior.

For (A) to be the case, our central moral-psychological capacities would be just those used in providing verbal answers to hypothetical scenarios; all distinct capacities would be, at best, peripheral moral-psychological capacities. But this is very dubious, even by the standards of theories that rely on studies of our abilities to draw the moral/conventional distinction. After all, the scenarios used in such studies concern the evaluation of *actions*. By independent standards, the production of actions shows up as a centrally important topic. Introductory courses in moral philosophy routinely characterize moral philosophy as the study of theories of right and wrong conduct. Such theories, and everyday moral questions, concern what people should do—i.e., what actions they should produce. Thus, the notion that action-production mechanisms are not central to moral psychology should be treated as deeply dubious.

So far as I can tell, (B) is false. I know of no work suggesting that variations found in abilities to draw the moral/conventional distinction predict, for example, variations in performance in Milgram-type scenarios. On the

contrary, the assumption of studies of the moral/conventional distinction, if applied to behavior, seems intellectualistic, in Ryle's sense: instead of the production of action being significantly tied to variations in context, the assumption of would-be extensions of moral/conventional testing to behavior is that action is produced by capacities that rely on information to which agents have conscious, first-person access before the fact, in abstraction from contexts that call for response. That is, first there is thought, then there is action. The studies performed in the course of the person-situation debate call this into question: people behaved in ways that surprised the agents themselves, in ways that were not predictable beforehand, and via mechanisms that seemed not to draw exclusively on information available to agents from a first-person subjective perspective.

Let's turn to the methodological thesis. If, as seems plausible, action-production mechanisms are of central importance to moral psychology, then the situationist tradition of empirical studies of action calls the methodological thesis into doubt. What is revealed by this tradition is the context-sensitivity of our action-production capacities. If this is at all correct, then it is plainly false that our central moral-psychological capacities can be studied in abstraction from interaction with actual people in actual contexts. Context-sensitivity is not susceptible to context-free examination.

Elliot Turiel has been the most explicit about defending the intellectualism and individualism of the reliance on moral/conventional studies. On page 8 of his landmark 1983 statement of Social Domain Theory, he claims that a methodological assumption of this position is that agents define, interpret, and judge social relations. This is a version of Rylean intellectualism. The implication is an assumption that when agents consider hypothetical scenarios in relaxed conditions, they are doing essentially the same thing as when they deal with actual agents and actual actions in real contexts that call for real-time responses: putting everything through a process of defining, judging, interpreting, and responding. At least partly on the basis of this assumption, Turiel explicitly rejects a division between natural and experimental contexts for the purposes of doing research on the structure of moral thought (ibid., 22). Such rejection amounts to a version of individualism: the experimental context of Social Domain Theory consists primarily in the administration of tests of the moral/conventional distinction. These tests, as we have seen, assess moral-psychological abilities in

abstraction from interactions with other agents in real contexts. Turiel explains the results of the Milgram studies in terms of the coordination of different domains of social knowledge—domains that contain correlatively different goals (ibid., 193–210). Perhaps most importantly, he relies on a 1980 review of the literature by Augusto Blasi, claiming that "it can be concluded . . . that an empirical relation exists between measures of moral judgment and measures of moral behavior" (193).

Let's examine how these considerations might work as responses to the line of thought presented here. The interpretation of the Milgram results is not implausible. Presumably this interpretation applies to the results of other situationist studies. But since this interpretation consists in an application of Social Domain Theory, its credentials have to be earned on the basis of empirical demonstration of Social Domain Theory. Such demonstration relies centrally on studies of our abilities to draw the moral/conventional distinction. Thus, the interpretation of Milgram does not count as independent evidence supporting Social Domain Theory and defending it against the questions being presently raised about its foundations. Instead, the interpretation is a consequence of this theory, these foundations, and these assumptions.

The same goes for the methodological thesis. Such assumptions have to be vindicated by the body of work performed with them as a basis. They do not count as data, independent or otherwise, in favor of such bodies of work. If there are findings that call such methodological assumptions into question, the assumptions lose credibility. This seems to be the case with Social Domain Theory and situationist psychology: Social Domain Theory might assume that agents perform such interpretive processes, but the findings of the situationist tradition call into question the extent to which the information accessible in explicit thought enters processes of action production. This intellectualist aspect of Social Domain Theory is particularly dubious in view of the surprising nature of the situationist results. It is the context-sensitivity of action-production demonstrated in the person-situation debate that calls into question the assumed individualism found in the rejection of the distinction between natural and experimental contexts in the Social Domain tradition.

The remaining plank in Turiel's defense of the intellectualism and individualism of the reliance on studies of the moral/conventional distinction is the 1980 review of the literature in which Blasi surveys a body of

empirical studies independent of those on which the Social Domain tradition rests. Although Turiel notes that Blasi's review finds consistency between thought and behavior for some topics but not for others (1983, 191–192), he is inclined to see the lesson of the review as supportive of the assumptions driving Social Domain Theory. However, the details are a bit more complex. Although Blasi begins his discussion of his findings by saying that there is "considerable support for the hypothesis that moral reasoning and moral action are statistically related" (1980, 37), he immediately qualifies this. There is much support for *specific kinds* of moral reasoning, including the idea that delinquents and non-delinquents have differences in moral reasoning, and for the idea that individuals at some higher stages of moral reasoning, in a Kohlbergian sense, exhibit relatively greater resistance to conform their judgments with others' under social pressure. But there is little support for other ideas, such as that "individuals of the postconventional level resist more than others the social pressure to conform in their moral action" (ibid., 37). In other words, the findings are mixed. Mixed results do not provide a firm basis for simple ideas about the relation between thought and action such as that exhibited in the explicit statements of assumptions offered by Turiel. Blasi is quite clear about this: "What was not learned in reviewing these studies, the successful as well as the unsuccessful, is the psychological meaning of significant statistical correlations between moral reasoning and action." (40) That is, his review does not support the claim that there is clear empirical support for the idea that individuals produce actions by going through processes of interpretation first.[5]

Daniel Wegner's more recent review of studies about the production of action in *The Illusion of Conscious Will* (2002) provides a useful perspective on these issues. On the basis of examinations of studies of normal action, of automatism (i.e., the performance of action without the experience of agency), and of the experience of agency without action, among other things, Wegner argues that our first-person experience of "conscious will" is produced by mechanisms that are psychologically distinct from those that produce action. He claims, quite explicitly, that "the actual causal paths [from causes to action] are not present in the person's consciousness" (68). Nevertheless, Wegner allows that there may be some sort of connection between the actual causes of action and what our experience of agency tells us. On the simplest interpretation of this idea, thought and action

may well turn out to be well correlated because they are *independent effects* of a common cause, not because first-person thought contributes to the causal process that produces actions. This sort of view provides an explanation of why the situationist results are so surprising: it's because our first-person experience of action is at least one step removed from the actual processes that produce actions.[6]

The present remarks do not require that Wegner have delivered the truth about how actions are produced. All that is required is an empirical basis for reasonable doubt about the methodological claims offered by Turiel. The combination of the closer look at Blasi's conclusions and at Wegner's more recent work constitutes exactly this basis. No convincing reply to the present doubts about the two psychological theses can be constructed from this part of Turiel's work.[7]

5.5 Obstacles to Philosophical Revival of the Traditional View

Let's look at some philosophical responses to the situationist challenge to virtue.

First, some philosophers—among them Athanassoulis (1999, 217–218) and Kupperman (2001, 242–243)—think that the situationist tradition shows the rarity of virtue, not its nonexistence. People are flawed, and virtue is difficult, so inconsistency in behavior is to be expected. This is in keeping with the dominant view of virtue stretching from Aristotle to twenty-first-century philosophers. However, this answer risks misunderstanding the nature of psychological inquiry, or at least its relation to traditional philosophical interests, and hence the burden of argument. In general, empirical science is one way of doing *ontology*—i.e., of developing an account of what really exists. To do this in a principled way, one needs a principled way of getting things into one's ontology. One cannot assume that things of a certain sort are in there; instead, one must bracket assumptions about existence, then let things into one's ontology as they pass the various tests that seem reasonable for such philosophical endeavors. Psychology is, in part, a principled way of cataloguing real psychological phenomena and the relations between them. Thus, the bracketing method applies to assumptions about psychology. From this perspective, the reply that virtue is rare puts the cart before the horse—it assumes exactly what has to be demonstrated and what should have been bracketed, which is

that character traits such as virtues and vices really exist. At issue are the viability and the grounds of exactly this sort of claim. This response risks being akin to criticizing children for not having adult psychological competencies, or to criticizing people for not flying like birds—we may not be constituted in such a way that makes these criticisms apt. Whether we are is exactly what has to be shown.

This is a rather general point, so let's apply it to the more specific case of the person-situation debate. Christian Miller (2003) places the Isen-Levin dime study in a wider context of attempts to replicate the results. Overall, replication was achieved in some studies (clearly in Levin and Isen 1975, less clearly in Batson et al. 1979) and not in others (Blevins and Murphy 1974; Weyant and Clark 1977). If the dime studies were the starting point for consideration of the contribution of situational factors to the production of behavior (and it can seem that it is the starting point, given Doris's emphasis on the original Isen-Levin study), then this wider context should give us pause. There seems to be little ground here for skepticism about the causal efficacy of such traits as generosity or helpfulness. But the Isen-Levin study is not the starting point, either historically or thematically, for consideration of situational effects on behavior. Conducted in the early 1970s, this study took place in the shadow of Milgram, and, more importantly, in a psychological context in which Mischel's 1968 review of decades of research on personality and behavior was fresh. Mischel's review earned for character traits the kind of bracketing that I just discussed abstractly. To adapt a term from bioethics literature on research ethics, Mischel's review put the psychological community into a position of theoretical equipoise (Freedman 1987): decades of research revealed that psychologists just did not know what the mechanisms that produce behavior were. At the time, whether character traits were significant sources of action was an open question. Against this background, the wider context of the dime studies reported by Miller looks different. The varied pattern of replication serves to maintain the theoretical equipoise of the time; that is, it leaves the question of the sources of action still open. There is no substantial comfort here for defenders of virtue. In fact, given the *prima facie* appeal many see in virtue-theoretical psychology, such a mixed pattern is a sobering reminder of how little empirical support this sort of approach has.

Suppose that virtue *is* rare because it is characteristic of abnormal psychology. What then? Suppose that cross-situational consistency of

behavior is a mark of saintliness or deep evil. This answer complicates many of the traditional interests of virtue theorists. For instance, in many of its forms, what is now known as virtue ethics assumes that the virtues provide an apt yardstick for the evaluation of human conduct. Some virtue ethicists think moral education should be aimed at inculcating the virtues and eliminating vices. (For examples and discussion, see Doris 2002, 121–127.) But if virtue is so difficult, perhaps it sets too high a standard for moral education and evaluation. If normal psychologies are not made up of traits such as virtues and vices, then perhaps moral education designed around these notions will be impractical. If the standard of moral evaluation is supposed to apply to normal agents, then perhaps a moral standard modeled after the psychology of abnormal agents is unfair. To adapt a rule of thumb from the legal studies, extreme cases make bad law. If the virtues and vices are characteristic of extreme psychologies, not normal ones, then perhaps philosophical theories based on these notions lose their *prima facie* plausibility.

Some theorists may choose instrumentalism about virtues and vices if situationist psychology turns out to be true (C. Miller 2003). The idea would be that even if the psychology implicit in the traditional view is not literally true of us, at least for everyday purposes these notions seem to make sense. Perhaps we can continue to use them as a kind of shortcut. This may be possible, but it will require modification or rejection of extant appeals to the virtues that treat them as real possibilities for humans. Moreover, the questions just raised about the fairness of virtue-theoretic moral evaluation and the practicality of virtue-theoretic moral education still apply, perhaps even with greater force. Finally, there is the risk of misunderstanding: insofar as the central concepts of virtue theory are not literally true, their use poses a serious risk of misleading people about the nature of human psychology.

Another type of response would be to give up the traditional view of character traits, yet to try to resuscitate psychologically viable notions of virtues and vices (Merritt 2000; C. Miller 2003). This would amount to a deflation of virtue such that it no longer centrally accounts for an agent's behavior and does not operate independent of the features of an agent's context. Assuming the vindication of situationism, this would be an empirically honest route to take, but it still faces problems. Like the instrumentalist strategy, deflationism poses the risk of misunderstanding. In particular,

there is the possibility of sneaking in reinflated virtue-theoretic notions. Insofar as people are accustomed to the traditional view of character traits, to use the same language for deflated notions is to ask for misunderstanding.

Finally, there is the most obvious strategy: just wait and see how psychologists resolve the empirical debate (Merritt 2000; Sreenivasan 2002; C. Miller 2003). This strategy exhibits more empirical humility than any of the others, and it may indeed turn out that situationism is false. But a victory for personality psychology would not automatically vindicate the notions of virtue and vice. For a resolution of the person-situation debate to be of direct aid to virtue theorists, it must be resolved *in the right way*. For instance, here are some purported features of virtues that have been particularly important to the interests of philosophers:

A. They produce good/right behavior;

B. They are sensitive to moral properties of states of affairs;

C. Their exercise is partly constitutive of rationality;

D. Their exercise is either constitutive of or conducive to individual well-being.

The person-situation debate has revolved around (A) only; to date it has had nothing direct to say about (B)–(D). With regard to (A), the psychological debate has concerned only the production of behavior, not its moral valence. It is quite possible that the kind of personality psychology that would emerge from the person-situation debate as an apt picture of human action-producing mechanisms would not vindicate the traditional view of these things in the least. Just as much as situationism, ongoing developments in personality psychology could be virtue's demise.

5.6 A Psychological Response

John Sabini and Maury Silver (2005) have argued that the challenge posed by situationist research in social psychology is not nearly as great as Doris and Harman have made out to be. I used Sabini and Silver's ideas in chapter 2, but in a different context than the one they intended. It is time to examine their response to the situationist challenge directly.

Sabini and Silver think that the important lesson of situationism is not that morally relevant behavior is deeply susceptible to environmental

influence, but rather that agents' behavior is "strongly influenced by what they take to be other people's perceptions" of the agent and the agent's context (2005, 559). Let's look at this position in some detail, partly because it is intrinsically important and partly because it raises issues central to the misunderstood aspects of situationism.

Instead of numerous variables that realize contextual sensitivity, Sabini and Silver argue that the effects on behavior are brought about through the nuances of peculiarly *social* interaction. To account for Milgram-type results, they offer as social pressures embarrassment and confusion brought on by the prospect of behaving in ways that show that one sees the world in a way different from others in the same situation (554–559). Their overall assessment of situationist social psychology is that it tells us something about moral behavior, but that its implications are not nearly as deep and revisionary as they are taken to be by such recent philosophical proponents as Harman and Doris.

Sabini and Silver's claim is that very particular kinds of social pressure affect behavior, which entails that this interpretation is not applicable to situations in which the particular social pressures are absent. Presumably there will be many such situations—many normal ones, given the surprising nature of the Milgram-type results. Thus, if Sabini and Silver are correct, action-production mechanisms will not be subject to surprising situational influence in many cases.

Since the interpretation of the Milgram-type results offered by Sabini and Silver invokes a very specific kind of social interaction and its effects, this interpretation does not apply in their absence. Specifically, the Sabini-Silver account applies only when embarrassment and confusion due to difference from others are psychologically relevant. Let's call this condition ECD. If the tradition of research for which this interpretation is offered includes studies to which ECD does not apply, Sabini and Silver's interpretation cannot be an adequate explanation of the results delivered by this tradition. At the very least they will have failed to account for studies that elude the scope of ECD. If other situationist mechanisms are required to account for situations that outside the ECD domain, they may also apply to situations to which Sabini and Silver's interpretation applies. That is, Sabini and Silver's explanation of Milgram-type results in terms of embarrassment and confusion may turn out to be at most a partial explanation of these results.

In fact, the situationist tradition has long contained studies that fall outside of the scope of ECD. Hartshorne and May 1928, which stands at the beginning of this tradition of research, is a good example. Two versions of cheating that Hartshorne and May tested were "The Copying Technique" (book 1, 49)—i.e., one student copying from another—and "The Duplicating Technique" (ibid., 51)—i.e., a student copying answers from an answer key. Performance of these techniques does not call into question whether one's view of the world differs from others'; it does not even require social interaction at all. These versions of cheating were part of the "IER" tests administered by Hartshorne and May. The average correlation between single tests of IER behaviors was 0.696, which Hartshorne and May claim is not high enough to enable prediction of deceptive behavior in one type of situation (e.g., a copying situation) on the basis of a score for another type of situation (e.g., a duplicating situation) (book 2, 213–213). It is worth noting that Hartshorne and May explicitly claim that their work shows that deceptive behavior is produced not solely by features of social interaction, but also by features of the specific situation and conduct the agent is considering and by the nature of the agent's relations to this situation and conduct (book 1, 397). This is in direct contrast to Sabini and Silver's interpretation of Milgram-type results in terms of the effects of specific kinds of social interaction.

The dime studies performed by Isen and Levin (1972) provide another important example. Besides their intrinsic interest, these studies are important because Doris discusses them at length (1998, 2002), whereas Sabini and Silver dismiss them very quickly (2005, 539–540). The social interaction involved in this study is so minimal—perhaps a few words or a glance exchanged with the person who dropped the papers—that ECD all but fails to apply here. This means that Sabini and Silver's interpretation of the Milgram-type results does not clearly apply to Isen and Levin's dime study.

Similar considerations hold for Isen and Levin's cookie study, in which unsolicited subjects were asked for some minor help by an experimental confederate. Some of the subjects had been given a cookie by a third party before this; others had not received a cookie. Isen and Levin found a significant difference in willingness to help based on whether or not a subject had received a cookie. Although this study involves social interaction, the specific nature of the interaction does not raise the issue of whether the

subjects' view of the world accorded with that of others. Consequently, the cookie study eludes ECD.

Since Sabini and Silver dismiss the aforementioned studies as not worthy of serious consideration, it is reasonable to think that they do not intend their account to apply to them.[8] This is a robust consideration only if Sabini and Silver's reasons for ignoring these studies are compelling. They offer two such reasons: (1) "[W]e just do not believe that picking up or not picking up your papers is a very important manifestation of a moral trait." (2005, 539–540) (2) "[B]eing in a bad mood . . . is the sort of thing that excuses the failure to notice some dropped pencils!" (2005, 540) Reason 1 is self-explanatory, but reason 2 requires comment. In the paragraph in which this statement is made, Sabini and Silver are discussing the effects of mood on attention. They assume that being in a good mood increases the attention one pays to one's surroundings, whereas being in a bad mood decreases it. Variations in behavior should be expected given such variations in attentiveness. This, presumably, is taken by Sabini and Silver to explain the variations reported by Isen and Levin. Neither of these reasons for dismissing the dime studies is compelling. Let's take them in reverse order. Reason 2 would be compelling if there were reason to think that all the non-helpers were in a bad mood upon leaving the phone booth. But there is no reason to think this. Sabini and Silver limit the options to being in a good mood and being in a bad mood, but this is a false dichotomy. I take it to be uncontroversial that there is a range of possible moods, many of them amounting to indifference rather than being positively or negatively valenced. Moreover, Sabini and Silver seem also to assume that not finding a dime that one wasn't expecting to find would suffice to put one in a bad mood. This is clearly false. The reasonable assessment of subjects' moods is that they fell across the whole range, and that *not* finding a dime had *no* effect on them. But this raises doubts about the variation of attention due to mood. If subjects' antecedent moods varied, and if this affected behavior via effects on subjects' levels of attentiveness to the world, then the reasonable prediction would be a mix of helping and non-helping behaviors in the situation in which no dime was found. But this is not what happened—there was an overwhelming absence of helping, compared with an overwhelming performance of helping in the situations in which a dime was found. Sabini and Silver's

interpretation of these findings in terms of attentiveness is plainly implausible.

Reason 1 brings up more general and important issues. I agree that whether or not one helps in such a situation is relatively morally insignificant. But it is precisely this moral insignificance that makes these studies important to consider. Doris makes this point (2002, 29–30): when assessing the empirical evidence for certain sorts of character traits, it is wrong to focus only on dramatic cases ("heroic" ones, to use Doris's term). Nobody is surprised by widespread failure to exhibit heroic goodness or dramatic evil. Very important and telling data come from studies of relatively ordinary and low levels of goodness and evil—exactly what is examined in the dime studies.

Moreover, Sabini and Silver's remark exhibits the general view of situationism as a psychological theory of specifically *moral* import. This is a telling point. There is reason to think that the challenge of situationism is more general than this—that it applies to philosophical considerations of the production of action in general, not only to a specific subset of actions with a certain level of moral value. If this is correct, then the moral importance of the behavior in question is neither here nor there: the empirical findings of situationist psychology apply to *all* action, whether morally significant or insignificant.

I do not doubt that the type of explanation of the Milgram-type results offered by Sabini and Silver advances important considerations. But we have good reason to think that it is, at most, a partial explanation. First and foremost, as we have seen, it does not apply to conduct in situations that do not raise the prospect of the right kind of embarrassment and confusion in social interaction. But its limitations are greater than this. If other mechanisms are needed to explain the Hartshorne-May results or the Isen-Levin results, then these mechanisms may well operate in cases to which the Sabini-Silver interpretation applies. That is, even where it applies, this interpretation of the Milgram-type results may well be a partial, incomplete account.

5.7 Revisiting the Situationist Challenge—The Production of Action

The philosophical adaptation of the person-situation debate has focused on explicitly moral behavior and on personality structures described in

explicitly moral terms. However, this way of casting the issues deserves a second thought. Is "moral behavior" really a subset of a bigger class of behavior, or is all behavior morally relevant? I am inclined to think that all behavior is morally relevant, and I shall argue that this is so. The implication of this is that the qualifier "moral" is meaningless as applied to the person-situation debate: *all* behavior should be thought of as being produced by context-relative mechanisms.[9]

As a first consideration, let's take another look at the sorts of studies performed in the person-situation debate. Consider two features of the Milgram studies:

(A) They examined *harming* behavior—i.e., *prima facie* bad, even impermissible action.

(B) They were performed using an *artificial* context—one whose characteristics were devised by and hence controlled by the experimenters—as opposed to a naturalistic, real-world context.[10]

Now consider the Isen-Levin cookie and dime studies:

(C) They examined *helping* behavior—i.e., *prima facie* good, even right action.

(D) They were performed in a naturalistic, real-world context.

Other studies exemplify the other possible combinations of these variables.

The combination of (C) and (B)—helping behavior in artificial contexts—is the topic of studies by John Darley and Bibb Latané. In one well-known study, subjects worked on questionnaires either alone or in small groups. Smoke would emerge from a wall vent. The question was the degree to which, if at all, being in a group affected one's inclination to go and find help on account of the "emergency." Seventy-five percent of subjects working alone sought help, but only 38 percent of subjects working with two other subjects left their questionnaires, and only 10 percent of subjects working with impassionate confederates of the experimenter went for help (Darley and Latané 1968; Ross and Nisbett 1991, 42).

Let's turn to the combination of (A) and (D)—harming or wrong actions in naturalistic, real-world contexts. For obvious reasons, this is a tricky combination to investigate. However, it is not difficult to find examples of real-world omissions of good, even right, behavior in the situationist

literature. The dime studies are simultaneously about such omissions. So is the well-known "from Jerusalem to Jericho" study of Darley and Batson (1973). In this study, students of the Princeton Theological Seminary were led to a building in which they were to give a lecture. On the way, they passed someone slumped in a doorway. The only variable that was examined that made any traceable difference to the likelihood of the students stopping to help was the degree of hurry to which they were subjected by the experimental protocol. Only 10 percent of hurried students helped, whereas 63 percent of unhurried students stopped to help the person in need (Ross and Nisbett 1991, 49). As for real-world harming, rather than artificial experiments, Doris offers twentieth-century examinations of genocide as providing the relevant data (2002, 53–61). The experiences of people in Nazi Germany and in early-1990s Rwanda offer much the same information as the Milgram studies—ordinary people in (slightly to radically) odd circumstances will participate in the murder of their neighbors.

We would have a serious reason to doubt the scope of application of the findings of the person-situation debate if the studies were clearly limited to a subdomain of either explicitly moral behavior or experimental context. But this body of work is not limited in that way—we have good reason to think that all kinds of moral valence and all kinds of experimental context have been addressed. If this tradition was not exhaustive in that way, then the claim that its findings could be extended from morally relevant behavior to all behavior would be under-supported. That, however, is not the case.

The crucial move is from morally relevant behavior to all behavior. To assess this case, let's turn from psychology to normative theory. Take consequentialism. Broadly put, consequentialists argue that the rightness or wrongness of an action depends on its consequences. For utilitarian consequentialists, the relevant kind of consequence on which to focus is happiness. According to Mill, "'the greatest happiness principle' holds that actions are right in proportion as they tend to produce happiness; wrong as they tend to produce the reverse of happiness" (1863, 7 in 2001 reprint). This implies that any action is morally relevant, since moral relevance is determined only by the nature of the consequences of the action and the available alternatives. These considerations arise with all actions. Thus, from the perspective of one central normative theory, the set of all actions is co-extensive with the set of morally relevant actions.

One may think that a broadly Kantian perspective implies something different. For instance, Kant claims in the *Grounding for the Metaphysics of Morals* that actions can be in accordance with duty yet can lack moral worth as a result of lack of motivation by duty: "[T]o preserve one's life is a duty; and, furthermore, everyone has also an immediate inclination to do so. But on this account the often anxious care taken by most men for it has no intrinsic worth, and the maxim of their action has no moral content. They preserve their lives, to be sure, in accordance with duty, but not from duty." (1785, A 398)[11] This could be interpreted as implying that some actions fall outside the moral domain owing to their psychological source. However, at least some Kantians reject this view. Onora O'Neill (1986) rejects it because she thinks Kantian deontology should be applicable to the conduct of certain sorts of groups. This conduct need not have a univocal psychological source from which it derives its moral worth. Instead of a particular sort of intention, O'Neill argues that the Kantian schema applies to the underlying principle of an action, which is present whenever action occurs. On this interpretation of Kant, all action turns out to be morally relevant (ibid.).

Similar considerations apply to virtue ethics, which brings us back to the person-situation debate. One variety of virtue ethics has at its core the idea that right action is action produced by the virtues, or by the best available motivation. (See, e.g., Slote 1995, 2001.) Something like this may be what springs to mind first when virtue ethics is mentioned. For instance, this is the way Walter Glannon casts virtue ethics as a whole in his recent introduction to bioethics (2004, 13–14). Michael Slote calls this sort of virtue ethics "agent-based" because it holds that "the moral or ethical status of acts is entirely derivative from independent aretaic (as opposed to deontic) ethical characterizations of motives, character traits, or individuals" (2001, 5). Slote's particular variety of this position sees benevolence as the most important virtue. Consequently, he claims that "an act is morally acceptable if and only if it comes from good or virtuous motivation involving benevolence or caring (about the well-being of others) or at least doesn't come from bad or inferior motivation involving malice or indifference to humanity" (2001, 38). Tom Hurka (2001, 223) interprets Slote's position as identifying *right* actions as those done from virtuous motives. Both Slote and Hurka see the nineteenth-century moral theorist James Martineau as offering a variety of agent-based virtue ethics.

According to Hurka, Martineau held that right action is action done from the most virtuous motives a person has to act upon (2001, 223). One way of developing such a position would entail that some actions are outside of the moral domain because of the sort of motive from which they spring. This idea can be usefully connected to a common and important response to the situationist position: that it overlooks the significance of the agent's perspective on the situations and behavior in question (an idea that was discussed in chapter 2 of the present volume). Both philosophers (e.g., C. Miller 2003; Sreenivasan 2002; Kupperman 2001)[12] and psychologists (e.g., Mischel 1999, Mischel and Shoda 1995; for discussion, see Doris 2002, 76–85) make this response to the situationists. Indeed, Ross and Nisbett (1991, chapter 3) discuss the importance of the agent's perspective as a commitment of social psychology just as important as situationism. In typical studies undertaken to show the importance of situational factors in producing behavior, the relevant descriptions of the situation and the subjects' action are provided by the experimenters, not by the subjects. These descriptions are, hence, objective (Ross and Nisbett 1991, 11; Sreenivasan 2002, 50) or nominal (Doris 2002, 76). Descriptions provided by subjects are called subjective or psychological. The response to the situationists claims that it is the subjective construal of the state of affairs and the agent's action that is relevant to the psychologists' purposes. The general idea is, at least partly, as follows: Situational factors do not directly produce behavior. Instead, cognitive mechanisms do. The situational information must somehow be received, interpreted, and used by an agent's cognitive mechanisms. At least some of these mechanisms will be ones that realize or are controlled by the agent's conscious understanding of the situation and of what sort of action it calls for. For example, Sreenivasan (2002, 65) cites a study by Charles Lord showing that cross-situational conscientiousness is much more consistent when individuals themselves view situations as similar.

Agents' understandings of their actions must be handled carefully. It is reasonable to think that, for psychological purposes, what matters is the agent's own understanding of what action is called for. But the present issue is one of moral theory, not psychology. For moral purposes, it seems not to be the case that one's perspective on one's own behavior is all that matters. Doris (2002, 80) gives this point some extended consideration in his example of mountain-climbers and "aipassion" (altitude-indexed

compassion). Imagine a person who exhibits classically compassionate behavior below 8,000 meters and classically incompassionate behavior above that altitude. From a nominal perspective there is inconsistency in behavior, but from the climber's subjective perspective this is a consistent display of aipassion. Here is Doris on this case:

> Suppose we castigate our aipassionate alpinist for what may plausibly be regarded as failures of compassion. He might reply that while he is inconsistently compassionate, he is quite consistently aipassionate. But the outrage and consternation observers may feel at his inconsistent compassion is unlikely to be assuaged by noting his consistent aipassion. Changing the subject is not an excuse. . . . (2002, 84)

In short, for moral purposes, it is not plausible to think that all that matters is the subject's subjective point of view. This extends not only to the matter of whether behavior is excusable or not, but also to the question of whether it is morally relevant in the first place. Indeed, that is how Doris's discussion begins. The example is derived from a comment by Eisuke Shigekawa that leads Doris to imagine peaks above 8,000 meters as "morality free zones" (Doris 2002, 78). Overall, the issue of the moral relevance of actions seems to hinge on whether moral standards can be reasonably applied to them. This issue seems not to be settled by the ways agents think of their own conduct. Hence, this line of thought—deriving from considerations of the psychology of virtue—does not undermine the idea that all action is morally relevant.

Suppose it is reasonable to think that all behavior is morally relevant. This means that the lessons of the person-situation debate apply not to moral behavior as some subset of behavior in general, but to the production of all action. We can sharpen this point using the personological assumptions examined earlier in connection with The Traditional View. The familiar view of the situationist challenge is that it applies to the production of behavior by character traits. Since questions of character arise in explicitly moral considerations, these assumptions are given the subscript *Moral* here:

Regularity$_{Moral}$ Character traits, which are dispositions to produce behavior, function in regular, patterned ways.

Autonomy$_{Moral}$ Character traits, which are dispositions to produce behavior, operate with substantial independence from contextual contingencies.

My claim is that the situationist challenge applies to a more general philosophical picture of the production of action. The Traditional View is a morally specific version of this more general picture, which I shall call The General View. Accordingly, the more general picture of action production is committed to more general versions of these claims:

Regularity$_{General}$ Action-production mechanisms function in regular, patterned ways to produce behavior.

Autonomy$_{General}$ Action-production mechanisms operate with substantial independence from contextual contingencies.[13]

If all behavior is morally relevant, then the two-prong attack on the Traditional View also applies to the General View, as do the various objections and responses.

5.8 Radical Situationism: Implications for the Psychology of Action Production

If situationism is radical, not only does it apply to the production of morally relevant behavior; it applies to the production of all behavior. In philosophy, since Donald Davidson's 1963 paper "Actions, reasons, and causes," causalism about action explanations has been widely accepted (and perennially contested). Davidson argued that actions are to be explained in terms of their causation by combinations of beliefs and "pro-attitudes." Davidson's paper is now considered the classic modern defense of a broadly Humean account of the production of action. Michael Smith (2004, 155–158) has argued that the Humean view of the explanation and production of actions is fundamental, and that other sorts of explanations make sense only as supplements to the Humean causalist account. I will address the primacy of the Humean schema shortly. For present purposes, a different and quite specific aspect of Davidson's position is worth examining.

Davidson proposes an *a priori* constraint on the construction of rationalizations. Davidson emphasizes that explanations of action must reveal "something the agent saw, or thought he saw, in his action" (1963, 685). When we construct a successful explanation of an action and thereby satisfactorily display what attracted the agent to this line of behavior "the agent is shown in his role as Rational Animal" (ibid., 690). Consider a case

in which I have two reasons for acting, and I act from one of them. Suppose that my reasons for writing these words include the intrinsic pleasure of productive intellectual activity and also includes the fear of missing a deadline that is approaching. Further, suppose that I act for the pleasure, not out of the fear. That is to say, at the time of writing, I saw the pleasure of writing as a compelling aspect of my options for action, and I saw the unpleasantness of missing a deadline either not at all or as a much less significant feature of my situation. Compelling explanations of my action must connect it to my motivations in ways that make sense—i.e., that do justice to the patterns of significance to which I was sensitive when acting. Let's call this the Rational Animal Constraint.

It should be evident that Davidsonian causalism exhibits both assumptions of The General View. The Rational Animal Constraint requires that patterns in my behavior be explained in terms of my psychological states, not in terms of my environment; this is the Regularity assumption. Since rationalizations make reference only to the agent's psychological states, the assumption is that only they are centrally responsible for the agent's behavior, which is the Autonomy assumption. However, there is a third feature of Davidson's position, or at least of one reasonable way of developing it, that is also relevant to present purposes. Again, this arises with The Rational Animal Constraint. One way of interpreting this is as requiring that rationalizations make reference to thoughts of which the agent was conscious at the time of acting. The light in which a course of action is shown as favorable to an agent is the light of consciousness. Accordingly, we can identify a third assumption of one stream of the dominant philosophical account of action production, to be added to those of The General View:

Conscious Access$_{General}$ Action-production mechanisms are, in principle, accessible to agents introspectively from the first-person perspective.

The case against the Regularity and Autonomy assumptions has already been laid out. This applies to Davidsonian causalism just as much as to virtue psychology. It should also be clear how the situationist case raises problems for Conscious Access. A very notable feature is how surprising the situationist results are. Again, nobody predicted in advance that Milgram's subjects would behave as they did. Likewise, in general we do not tend to suspect high correlations between tiny good fortune, such as

finding a coin, and helping someone. This strongly suggests that our action-production mechanisms are not easily accessible through introspection; their workings have to be revealed through careful, objective experimentation. Recall the work presented by Daniel Wegner in *The Illusion of Conscious Will* (2002), briefly presented above. Wegner argues that our first-person experience of "conscious will" is produced by mechanisms that are psychologically distinct from those that produce action. He quite explicitly claims that "the actual causal paths [from causes to action] are not present in the person's consciousness" (2002, 68). Conscious Access faces serious empirically based challenges from situationist psychology and from wider-ranging psychological studies of the production of action.

The situationist challenge is radical in the sense that it applies to the very root of consideration of the production of actions. It is *not* radical in the sense of being altogether revolutionary. Situationism calls for revision of familiar philosophical schemas of the production of actions, such as Davidsonian causalism, but not for their wholesale rejection.

For one thing, situationism does not imply that beliefs, desires, and other pro-attitudes[14] do not produce actions, but it does imply that they *alone* do not produce actions. The reasonable way to account for widespread behavioral inconsistency is, barring complications that will be noted shortly, simply to add other things to the list of psychological contributors to the production of action.

The second implication of radical situationism stems from the findings of contextual sensitivity. Whatever items the true list of action-production mechanisms turns out to contain, they should be conceived of in a manner that accounts for their sensitivity to context. On this point, situationist psychology dovetails with the concerns of the individualism-externalism debate in philosophical psychology. Recall that individualists contend that mental properties are realized solely by intrinsic properties of agents. Externalists deny that such intrinsic properties do all the work, and recognize a role for relational properties in the realization base of mental properties. One way of interpreting the empirical findings of the situationist tradition is as indicating that action is produced by mechanisms realized by properties of both the agent and the agent's environment. Such a system is the topic of the next section.

5.9 Wide CAPS: Deep Externalism in Action

The over-arching purpose of this section is to provide the beginnings of a deeply externalist account of human action-production mechanisms. More specifically, I will argue that a wide interpretation of Yuichi Shoda and Walter Mischel's Cognitive-Affective-Personality-System (CAPS) account of the nature of the personality structures that produce behavior is plausible by the standards of the data and the general approach to psychological theorizing used by Shoda and Mischel. CAPS is an independently interesting and important model of the personality structures responsible for the production of action, but it is particularly important to examine in connection with latter-day debate about situationist psychology. Philosophers on both sides of this debate (Doris 2002; C. Miller 2003; Sreenivasan 2002, 66) have been attracted to the work of Shoda and Mischel. This is understandable in view of Mischel's role in generating the person-situation debate and Mischel and Shoda's declared aim to straddle the gap between the poles of this debate. CAPS derives in no small part from paying attention to both the role of features of an agent's environment and the contributions from an agent's psychology to the production of action.

Shoda and Mischel are officially silent on the individualism/externalism issue. However, in view of the structure of CAPS and the overwhelming tendency in psychology to frame individualistic hypotheses and explanations, it is reasonable to interpret CAPS along individualistic lines. Accordingly, for present purposes, I shall present CAPS as explicitly individualistic. I will call the deeply externalist alternative "Wide CAPS" to make explicit both its externalist aspect and its relations to the work of Mischel and Shoda.

Action-production is too vast a territory to address completely in a brief section. Accordingly, let's narrow our focus. The explanandum, for present purposes, includes two aspects of the findings of the person-situation debate: that individuals exhibit contextually sensitive variance in their behavior, and that the situational factors that elicit behavior can be surprisingly insignificant.

Individualistic Hypothesis—CAPS

Both of the aforementioned features of the person-situation debate are important to the work of Mischel and Shoda, who emphasize the

importance of individual patterns of variance as the hallmark of personality rather than cross-situational constancy (1995, 248–250). Indeed, Mischel and Shoda claim that their position is "intrinsically contextualized" (Shoda and Mischel 2000, 408). The system that is offered as an account of the source of behavior is "intrinsically interactive with the social world in which it is contextualized" (Shoda and Mischel 1996, 418). The systematic nature of the overall personality structure is offered to account for the sensitivity found in the Milgram-type studies.

Here is the CAPS account in more detail: Intra-individual context-sensitive patterns of variance in behavior offer psychologists a challenge. Regularities seem to call out for explanation by reference to constancy of some sort, but variance calls out for explanation in terms of differing factors. The explanans offered by Mischel and Shoda is a system of cognitive and affective units. These units include, but are not limited to, beliefs, plans, values, and "encodings" (Mischel and Shoda 1995, 253; 1996, 416; 2000, 420). This system is explicitly conceived of as located within the physical bounds of the agent; in diagrammatic representations of the CAPS, the environment is always located as outside the bounds of this system (1995, 254; 2000, 413). Besides the content of these units, differences in behavior between individuals are accounted for in terms of differences in their activation and, especially, in their organization (1995, 253). The complex yet systematic organization of this system accounts for the Milgram-style effects. Mischel and Shoda explicitly use connectionist networks as an image for the systematic complexity that they have in mind. A seemingly insignificant input to such a multiply connected system can have surprising, even unpredictable effects. Since the units are connected systematically, these effects are not random. Given nearly constant arrangement of the system, roughly the same input will produce roughly the same output. Mischel and Shoda call these stable patterns "if-then situation-behavior relations" (1995, 248–250): if psychologically salient situational factor A is present, then behavior X will be the result. These if-then relations codify intra-individual variance. Constant patterns of variance are accounted for by the structure of the system, which overall is slow to change.

To emphasize the individualistic aspect of CAPS: Convinced of the explanatory power of a system of mediating units, Mischel and Shoda assume that it is to be located within the physical bounds of the agent.

The issue then is what, exactly, are the units that constitute the system; this is the title of a section in Shoda and Mischel 1996. Shoda and Mischel answer in a pluralistic spirit, drawing on a variety of research programs. The kinds of items that are attributed to the individualistically construed system include, but are not limited to, consciously accessible psychological states such as beliefs and plans. Although Mischel and Shoda are highly cognizant of the importance of context, features of an agent's situation are not included in the personality system. Instead, they ultimately function only as a source of input to the system.

These points can be put in terms of the Regularity and Autonomy assumptions. Mischel and Shoda reject the Autonomy assumption, in that the psychological items attributed to individuals are not assumed to operate independent of features of the agent's context but are instead deeply integrated with the agent's context. However, Mischel and Shoda sit on the fence with regard to the Regularity assumption. They reject it insofar as agent-environment "if-then" relations are responsible for patterns in behavior, but they retain it insofar as the psychological items that are causally relevant to the production and explanation of behavior are attributed to individual agents.

Externalist Hypothesis—Wide CAPS

Mischel and Shoda's conviction about the explanatory power of a psychological system of mediating units is well founded. Intra-individual variance in behavior is explained by such a system in both CAPS and Wide CAPS. What is worth questioning is the assumption that this system must be located within the physical bounds of the agent. This is worth questioning by Mischel and Shoda's standards, since they characterize the system that is putatively responsible for the patterns of behavior production found in the person-situation debate as "intrinsically contextualized." A substantial way to inherently contextualize one's position, different from that pursued by Mischel and Shoda, is to reject the assumption that the cognitive system in question is located within the physical bounds of the individual. Instead, the intrinsic connections between agent and context, codified by Mischel and Shoda as if-then relations, should be taken as evidence of the causal and functional integration that is the hallmark of systemicity itself. That is, the system that is relevant to the explanation of the patterns of behavior found in the person-situation debate is a wide one in which individuals

play a role, rather than one located solely within the physical boundaries of individuals.

What psychological units are to be attributed to individuals under this hypothesis? Again we can follow Mischel and Shoda, but we must be cautious. Overall, this is an empirical issue. At this point, individualistic CAPS faces only one question: What is the empirically warranted way of explaining behavior? Wide CAPS faces this question too, but it also faces a second question: What features does an individual need to participate in the wide system that, by hypothesis, produces behavior? The cognitive and affective units that Mischel and Shoda adopt from other research programs are a good place to begin in answering these questions, as these have at least partly earned their way onto the scene through their explanatory efficacy. But at this point it is an open question whether these units should be attributed to individuals. In view of the overwhelming assumption in favor of individualism in psychology, the research programs on which Mischel and Shoda draw are, in all likelihood, individualistic ones. Wide hypotheses for their respective explananda ought also to be framed and tested. Since this has not been thoroughly done for psychology, we must be careful about connecting the ideas from these research programs to Wide CAPS. Here is how to proceed: In Wide CAPS, the psychological units deployed in other apparently successful research programs can be applied to either of two objects : the individual and the wide system. In principle, it is possible that the units attributed to the individual in Wide CAPS will be the same as those attributed to the individual in CAPS. However, it is also possible that the units that Mischel and Shoda adopt from other research programs should be attributed to the Wide CAPS system, not to the individual. What is attributed to the individual will depend on what is required for participation in a wide system that is more specifically characterized in terms of the units adopted from other research programs.

Here is another way to put this point: It is reasonable to interpret Mischel and Shoda's units—e.g., beliefs, desires, and encodings—in terms of information processing. Mischel and Shoda assume that the information-processing tasks relevant to the production of behavior take place within individuals. The units attributed to individuals are, by hypothesis, the units that accomplish the relevant information-processing tasks. In contrast, for Wide CAPS at least some information processing is performed between the individual and the environment, not within the individual.

The evidence that is taken by Mischel and Shoda as warranting attribution of an information-processing unit to an individual works differently for Wide CAPS: it warrants attribution of information-processing tasks either to an individual or to the wide system. Consider beliefs. One reason to attribute beliefs in an explanation of behavior is to account for the cognitive representation of certain features of the environment. CAPS assumes that processing of this sort of representation takes places within the physical boundaries of the individual. In contrast, Wide CAPS countenances the possibility that information processing of this kind takes place between the agent and the environment. If this is correct, then it is an open question whether beliefs need to be attributed to the individual in order to be part of the wide system. Overall, when evidence and explanatory need justify attribution of a task to the wide system, then we lose the grounds that Mischel and Shoda have for attributing a certain sort of psychological unit to an individual. Instead, we will likely have reason to attribute to the individual a different sort of psychological unit—one that facilitates participation in the relevant sort of widely systematic information processing. Maybe individuals will turn out to be characterized in terms of a wide array of beliefs and desires, and maybe not. Further evidence and testing of more refined hypotheses is required.

Under the individualistic hypothesis, the surprising nature of the Milgram-style results is explained as follows: Some of the psychological units of CAPS are not familiar, first-person-accessible sorts of psychological states. Thus, one is not necessarily in touch, from within, with the psychological units that produce one's behavior. This explanation is also available to Wide CAPS, but so is another explanation: If there are differences between first-person access to psychological processes that happen within the physical boundaries of one's skin and first-person access to psychological processes that are realized by a system in which one plays a role, such that wide processes are less accessible, Milgram-style surprises will be one result.[15]

Wide CAPS rejects both the Regularity assumption and the Autonomy assumption. The features of individuals that are causally responsible for behavior are not assumed to function independent of the agent's context. Nor are patterns in an agent's behavior assumed to be produced by psychological items attributed to the agent. Regularity and Autonomy might be empirically vindicated, but they are not assumed by Wide CAPS.

Reflections

Since CAPS and Wide CAPS have been devised using the same evidence and the same general approach to psychological theorizing, additional evidence and/or attention to the existing evidence is needed to decide conclusively between them. But it is worth asking whether there is any reason so far to prefer one over the other. This returns us to the question, addressed in chapter 1, of when externalist hypotheses are warranted. The answer to this question was that they are warranted when there is evidence of the causal and functional integration characteristic of systematic individual-environment relations. This gives us the beginnings of an answer to our new question. The greater the causal-functional integration between individual and environment with regard to a given psychological phenomenon, the more warrant there is for externalist hypotheses about that phenomenon. With regard to the findings about behavior at the core of the person-situation debate, we have reason to think that there is quite a high degree of individual-environment integration. Recall that Mischel and Shoda themselves characterize their position as *intrinsically* contextualized. That is, they see the data as requiring explanation of a sort that explicitly includes the agent's context. One issue that differentiates CAPS and Wide CAPS is how to provide such an explanation. The more seriously one takes claims of this sort, the more reason one has to pursue externalist hypotheses in this domain.[16]

To pursue this line of thought further, I shall speculate about some specific mechanisms characteristic of the Wide CAPS system.

5.10 Perception and Behavioral Inhibitors and Enablers

What features of an agent's environment could be included in the mechanisms responsible for the production of behavior? To address this, I will show that features of an agent's environment can figure constitutively, not merely as input, in either the belief or the pro-attitude aspect of a primary reason. This demonstrates that the fundamental aspects of a Humean approach to action need not be filled by intrinsic features of an agent. In pursuit of this, and to develop Wide CAPS, let's connect the findings of situationist psychology with externalist considerations of the most familiar way in which cognitive processes meet the world beyond the physical bounds of the agent: through perception.

There is well-developed externalist work on visual perception. Foremost in this tradition is the work of J. J. Gibson. However, as we saw in the preceding chapter, Susan Hurley has argued that more recent work in neuroscience calls for even more deeply externalist revisions to our ideas about perception than Gibson's work offers.

Mark Rowlands (1999, chapter 5; 2003, 169–173) has argued that a Gibsonian approach to perception is an externalistic one. On an individualistic approach to perception, perceptual mechanisms are constituted by intrinsic features of agents. The environment figures as a possibly mysterious trigger to perceptual processes that happen within the physical bounds of an agent's body. Crucially, all information processing happens within the physical bounds of the agent. In contrast, the Gibsonian approach emphasizes the psychological reality of the environment itself.

Rowlands presents the "optic array" as the core of Gibson's externalistic position on perception. The optic array is "an external information-bearing structure" (1999, 107; 2003, 171). Space is filled with rays of light reflecting from every surface. At every point, these rays converge. Because of this, "there is what can be regarded as a densely nested set of solid visual angles which are composed of inhomogeneities in the intensity of light" (1999, 107). The structure of the optic array is "nomically covariant" (1999, 108) with the structure of the environment—the structure of the environment determines the intensity and angle of the light at any given point—so an organism that can access the information contained in the optic array thereby gains information about the environment. On this account, there is no need for an agent to re-represent the information contained in the optic array within the physical bounds of the agent's body. Instead, some of the information processing involved in visual perception happens between the agent and the optic array. Perception is partly constituted by the informational resources of the environment.

Action is central to the Gibsonian account of visual perception. The optic array is sampled by an organism moving through its environment. However, action affects visual perception only by changing the input to visual perception mechanisms. For this reason, passive movement does as well as active movement for Gibsonian purposes; the degree and the kind of intentional control play no direct, constitutive role in visual experience for Gibson. Hurley argues that subsequent developments in neuroscience and the study of perception give us reason to doubt this aspect of the

Gibsonian picture. The modifications added by Hurley deliver an even more deeply externalist view of visual perception and the mind than the Gibsonian account.

The Gibsonian approach to visual perception gives the environment beyond the physical bounds of the agent a constitutive role. Hurley's modifications preserve this, but they integrate action with perception more deeply.

Recall the discussion of vertical and horizontal modularity in the preceding chapter. The classical sandwich view of the mind, which is designed around vertical modules, exemplifies two assumptions. According to the linear assumption, cognitive processing is one-way: from the world to perception to higher cognition and back to the world. According to the instrumental assumption, perception and action are not constitutively related, but are related only as means (Hurley 1998, 419). The Gibsonian view rejects the instrumental assumption and instead asserts that visual perception is partly constituted by action. But the insistence that action's contribution to perception is accomplished only by affecting the input to visual mechanisms effectively retains the linear assumption.

As we have seen, Hurley argues that this view of the mind, and specifically its vertical modularity, has been called into question by cognitive science and by neuroscience. (For a list of research programs offered in support of this claim, see Hurley 1998, 408). Instead of a mind composed of constitutively distinct vertical modules, Hurley argue, neuroscience reveals a mind featuring horizontal modules. There are two important features of the horizontally modular view of the mind presented by Hurley. First, each module is constituted by both input and output functions. Functioning within each module can include feedback from relatively more downstream to relatively more upstream stages of processing. Second, there is no layer that, by itself, constitutes higher cognitive functioning. Instead, this is something that emerges from the interplay of the specific perception-action layers.

Gibson allows only for instrumental content dependence of visual perception on action, in that action affects perception only by affecting input to perceptual modules. Hurley argues that studies in visual perception demonstrate non-instrumental content dependence. This is revealed by changes in perceptual content even when input to perceptual mechanisms is held constant. Since such variation cannot be explained in

terms of variation in input, feedback *within* the module from variation in other kinds of subpersonal processing to perception must be invoked instead.

For example, in now-familiar studies by Ivo Kohler, subjects wore goggles that were tinted in a systematic manner: the left sides were tinted blue and the right sides were tinted yellow. When subjects wearing these goggles looked left, the visual field appeared bluish; when they looked right, it looked yellow. After wearing the goggles for several weeks, subjects adjusted; the tinting tended to disappear. This means that the visual system adjusted for the systematic change in input. When the adjusted subjects removed the goggles, the visual field would appear yellow when they looked left, blue when they looked right. Hurley (1998, 287) quotes Kohler as saying that the eye motion signals the visual system to make the color adjustment. Hurley adapts this study into a thought experiment about non-instrumental content dependence. Imagine the adjusted subject wearing the goggles and looking at a uniform white field. Since the subject is adjusted, the visual field appears white. When we imagine the subject removing the goggles while looking at the uniform white field, we can expect that eye movements to the left would make the visual field appear yellow, and eye movements to the right would make it appear blue. Since the subject is looking at a uniform white field, input is constant. The variation in the content to visual perception is not due to variation in input. Instead, variation in output—in motor commands to the eyes to move left or right—seems to account for the change in visual content, in the absence of change of input. According to Hurley (1998, 289–292), this indicates the possibility of non-instrumental content dependence for visual perception. If this is correct, then the linear assumption of the classical sandwich view of the mind should also be given up.

Here is how all of this is connected to externalist concerns: The view of the mind that Hurley is resisting treats it as essentially an information processor. Information comes in at certain points, gets manipulated, then gets passed on or gets manipulated some more. Types of information processing are discretely isolated from each other. The kind of information processing most familiar to us—the processing characteristic of higher cognitive processes—is not only separated from the mechanics of other kinds of processing, but also separated from the world beyond the physical boundaries of the person. In contrast, horizontal modularity belongs to a

picture of the brain and mind as instruments for participation in this wide world.[17] Not only are the boundaries between perception, action, and higher cognition much more porous on this view, so are those between person and world. Hurley thinks that the processes by which, e.g., output functions affect input functions need not be contained within the physical boundaries of the organism (1998, 332). In short, vertical modularity lends itself to internalism about mental content and mental processes, and horizontal modularity lends itself to externalism about these things. Indeed, Hurley explicitly ties her discussion of horizontal modularity to a defense of externalism (1998, chapter 8 especially).

Overall, the work of Gibson and Hurley suggests a growing emphasis on the blending of perceptual and action-producing processes in which features of an agent's environment can play a constitutive role in cognitive processing. Let's suppose that something like this is a plausible account of at least some of what is going on in visual perception.[18] For this to be relevant to the production of action, the information in the optic array (to use the Gibsonian terminology) must be able to play the role of either the belief or the pro-attitude in the Davidsonian Humean schema. Long-standing work in the situationist tradition suggests that this can in fact be the case. Following Kurt Lewin, one topic of situationist research has been so-called channel factors, or ways in which aspects of an agent's environment either enable or inhibit action (Ross and Nisbett 1991, 10). Channel factors can be usefully reconstrued in terms of the Davidsonian Humean schema, thereby giving this schema an externalist twist.

Let's consider typical roles for beliefs and pro-attitudes in the production and the explanation of action. Suppose someone asks why Andrew went to the fridge, and the reply is "Because he wanted a Rochefort 10." We can fill this out to fill the Davidsonian Humean schema:

Pro-Attitude A desire for a Rochefort 10.

Beliefs That there was Rochefort 10 in the fridge, and that he could get Rochefort 10 by going to the fridge, etc.

In short, the pro-attitude supplies the agent's *end* and the beliefs provide information about *means* to that end.

We can interpret channel factors that enable action as external features that provide an agent with an end for action, with the means of achieving an end that the agent already has, or with both. The Milgram studies are

most aptly interpreted primarily in terms of the provision of ends. It is not the case that the subjects in general wanted to hurt people and that the experimental protocol gave them the means of accomplishing this. Records of subjects' experiences of distress give the lie to this interpretation (Doris 2002, 42–45). Instead, subjects' involvement in the studies gave them new ends, perhaps in connection with some already existing ends (such as to comply with the study) and certainly in conflict with already existing ends (such as not to hurt people). The Milgram studies also provided the means to accomplishing such ends, and it is here that the work in visual perception most aptly connects. The Milgram subjects' optic array contained information about the "shocks" they were applying, for instance.

An externalist position on visual perception and the provision of ends is attractive for the Darley-Latané smoke studies. Alone, the perception of smoke is the perception of an apparent emergency that the subject must address, because no one else can. When others are present, what is perceived is more complex—for example, a situation that might or might not be an emergency, and other people who might or might not deal with the emergency, or who might or might not have better insight into whether the smoke indicates an emergency. Of course, in both cases the information perceived can be combined with already existing and even individualistically realized ends. Externalism requires only that some aspects of an agent's psychology happen between the agent and the environment, not that all of it does.

Besides enabling action, channel factors also inhibit conduct. This is central to certain ways of interpreting the Milgram studies. Subjects with the ends of avoiding harming others and of ceasing participation in the study find themselves externally inhibited by, e.g., the firm encouragement of the supervisor, or by a physical layout of the study that provides easy means of "hurting" the learner but no easy means of terminating the session.[19] Various research programs suggest that inhibitory processes of one kind of another are a normal and important part of morally competent behavior. For example, Marc Hauser (2006) uses studies of non-human primate cooperation to explore the importance of the inhibiting of one sort of action in favor of others in pro-social conduct. Some models of psychopathy accord a central role to impairments of certain sorts of inhibitory systems (Fowles 1980, 1988; Kring and Bachorowski 1999; the integrated emotion systems model of reactive aggression of Blair et al. 2005 is

a development of their own earlier violence-inhibition model—2005, 79, 122–124). Combined with Wide CAPS, this line of thought suggests that external channel factors that can inhibit the production of action are, in principle, important aspects of action-production systems. They are also normally overlooked.

To access environmentally encoded information, the agent does not, in principle, require the cognitive resources to reproduce that information within the physical bounds of his or her body. Instead, what is required is some way of tracking and incorporating that information in information-processing systems, such as the ones for the production of action. In principle, the perception of means and the perception of ends could be realized by exactly the same items within the physical bounds of the individual agent. Recall the discussion of mirror neurons in chapter 2.

One might attempt a conceptual objection at this point. Michael Smith (1994, 111–112; 2004, 156) argues that beliefs and desires must be distinct states. Our idea of a belief is of a state that must fit the world, and that tends to go out of existence when it is discovered that it does not fit the world. In contrast, our conception of a desire is one of a state which the world must fit, and which does not tend to go out of existence when it is discovered that the world in fact does not fit it. Since these descriptions cannot be simultaneously filled by a single item, beliefs and desires must be psychologically distinct states. Daniel Goldstick (2006) has argued that this is merely a failure of imagination, and I am inclined to agree. Instead of distinct items, we should take the Davidsonian schema as identifying *analytically* distinct psychological jobs. Whether these jobs can be performed by a single state, or whether their psychological realizations must be physically distinct, is an *a posteriori* issue (as is whether they are contained within an organism's physical boundaries). Work in both psychology and philosophy suggests the psychological possibility of simultaneous performance of the belief and pro-attitude jobs by a single state.

Ruth Garrett Millikan (1996) gave the name "pushmi-pullyu representations" to states that both represent the world and direct the system that has such representations to perform an action. She argues that such hybrid representations should be seen as primitive, and that more specialized states that only represent the world or only direct the organism to do something—i.e., standard beliefs and desires—should be seen as subsequent developments characteristic of more complicated processing systems.

Non-human examples of pushmi-pullyu representations include many signals that animals use, including bird songs and bee dances (ibid., 146). Interestingly, Millikan draws examples of human pushmi-pullyu representations from moral discourse. The kind of sentence used in moral education, particularly of children, often has a pushmi-pullyu form (ibid., 153–155). The utterance "Grown-ups don't hit each other," directs one not to hit by representing a state of affairs.

For a more general philosophical discussion, let's turn again to Hurley. One of Hurley's principal concerns is that the classical sandwich picture of the mind unduly simplistically maps personal and subpersonal levels onto each other. Hence, distinctions found at the personal level that are described in terms of perceptual content and the content of intentions respectively are taken to be functions of distinctions in input and output respectively. Hurley argues in a variety of ways that both distinctions and invariants in personal-level content can be functions of relations between input and output. In such cases, there is no one-to-one mapping of personal-level distinctions to subpersonal-level distinctions. Hence the alternative view of the mind gives us horizontal modules constituted by feedback relations that cut across the subpersonal boundaries between input and output. Higher cognitive processes are presented as emerging from the interplay of horizontal modules, rather than depending on central subpersonal processes distinct from both input and output systems. Abstracting from considerations of the details, if Hurley is correct that distinctions in input can have a constitutive role in the formation of basic intentions (363–365, 389–400), then a single state is playing, simultaneously, both a world-representing role and an action-directing role.[20]

Something similar is found in the work on visual perception, and specifically the "visuomotor system," by the psychologists Melvyn Goodale and David Milner (1995). Their position emphasizes the independence of the cognitive processing responsible for perceptual experiences from the processing responsible for the production and control of action. This is a development of work on the functions of the dorsal and ventral streams of visual processing (Ungerleider and Mishkin 1982). According to Goodale and Milner, the ventral stream yields long-term perceptual representations (Goodale 2001, 192). The dorsal stream concerns "moment-to-moment information about the location and disposition of objects with respect to the hand or other effector being used and thereby mediate the visual

control of skilled action" (ibid., 192–193). Lesions that affect one stream but not the other yield dissociations of action from perceptual experience. Goodale relates the case of DF who suffered from brain damage due to carbon monoxide inhalation. DF could not recognize the faces of people she knew well; nor could she "identify the visual form of common objects" (197). Nevertheless, DF could use visual information to produce skilled movements, such as shaking hands or turning a door handle (199–203). For present purposes, note that dorsal stream states simultaneously represent specific aspects of the world and guide action.

Conclusion

Once again recall the rough notion of systemicity introduced in chapter 1:

_____ systems must be causally and functionally integrated chains of _____ resources, and these, individually and collectively, must play a replicable causal role in _____

The last few sections of the present chapter explored ways in which this schema might be filled out for perception and action. For perceptual processes, the optic array is arguably a central perceptual resource that plays a crucial and replicable causal role in perception. Notably, it is located outside the physical bounds of the agent. There is some reason to believe that it can also play a role in action production, and hence in filling out this schema for action-production systems. Certainly there is reason to take seriously the importance of perceptual processing as one of the diverse capacities that make up Wide CAPS (and CAPS itself, presumably).

All of this raises problems for the Davidsonian Rational Animal constraint. Arguably, CAPS provides one way of pursuing this aspect of the Davidsonian position. Wide CAPS, however, provides at least the possibility of a break with this aspect of Davidson's position. For Davidson, Mischel, and Shoda, the rationality of the agent is preserved by attributing the psychological causes of action to the agent. Wide CAPS presents the theoretical possibility that the causes of action should be attributed to a system of which the agent is only a part. This presents the further possibility that the first-person perspective of the agent is alienated from the causes of action in such a way that it no longer makes sense to ask what the agent

saw as attractive in the given line of conduct. If there are psychological obstacles to gaining first-person access to external aspects of perceptual and action-production processes, then it is not the case that explanations of action must preserve an agent's perspective of the situation and conduct. Davidson begins with the assumption that explanations of action must display the agent as a rational animal. This is not a conclusion for him, but the platform on which his position is built. In contrast, deeply externalist accounts of the production of action such as Wide CAPS present the possibility that the empirical details in this domain might not deliver this picture of agents. The possibility of deep externalism means that the rational animal model of action production must be argued for on an empirical basis rather than assumed as an *a priori* foundation for thought about action.

My aim in this chapter has been to demonstrate the possibility, the plausibility, and the moral-psychological importance of an externalist position about the production of action. Its possibility was evident, in general form, on the basis of the taxonomy of externalist positions presented in chapter 1. However, in this chapter I aimed to bring this abstract possibility closer to the ground by connecting CAPS and situationist psychology to each other and to externalist themes from philosophy. The plausibility of Wide CAPS as a general model for the production of action stems from the combination of the possibility of CAPS, a broadly externalist position on perception, and the interpretation of the findings of the situationist tradition in psychology in terms of the wide Davidsonian Humean schema. The moral-psychological importance of this position derives from the importance of the situationist tradition and its implications for the production of action as a central moral-psychological topic. This is a lot, much of it only suggestively handled here, but I trust that what has been presented suffices for giving heft to this aspect of the Wide Moral Systems Hypothesis.

6 Psychological Pluralism, Environmental Sensitivity, and the Bounds of Morality

I have made a case that cognitive systems that extend beyond individual agents into the wider world, including other agents, are important constituents of the psychology of normal moral agency. Along the way, I have made the case that the psychology of moral agency turns out to be, at virtually every turn, heterogeneous. In this chapter, I shall prioritize my topics differently. Pluralism will be my primary topic; externalism will be secondary.

Just how fragmented is the psychology of moral agency? There are two levels of abstraction at which to pose this question. At the higher level, the question to ask is "Just what capacities are characteristic of normal moral agency?" In chapters 2–5, I have examined moral judgment, moral reasoning, action production, and attributions of responsibility. Should anything be added to this list in order to paint a complete picture of the psychology of moral agency? At the lower level the issue is pluralism within each segment of the taxonomy of the essential features of moral psychology. I have argued for pluralism at this level in each of the four preceding chapters. As I add topics at the higher level of abstraction, I shall keep in mind the possibility of pluralism at this lower level.

I shall broach the question of the extent of the heterogeneity of moral psychology via an attempt to map the psychological contours of morality. I shall do this by creating a taxonomy of forms of amoralism. It is common psychological practice to use malfunctioning as a guide to normal functioning. My map of amoralism is constructed in this methodological spirit. For each way in which people are insensitive to moral demands, it is worth asking what psychological mechanisms are responsible for securing the appropriate sort of sensitivity. By so doing, we can construct, at least *prima facie*, a list of topics that must be addressed by a complete account of our

moral psychology. In the specific context, this means that we will have a tool to use to determine what should added to the Wide Moral Systems Hypothesis (WMSH) as it is developed.

The second topic of this chapter is closely linked to the first one. In chapters 2–5, I have presented a pluralistic view of the psychology of moral agency. The taxonomy of forms of amoralism expands this pluralism. The second topic is the question of what to make of this. What explanation might there be of the psychological pluralism of moral agency? My answer to this question returns externalism to the discussion. Briefly, my suggestion will be that our moral minds are psychologically heterogeneous in part because they are significantly widely realized and the world in which we operate is heterogeneous.

The third topic of this chapter is the practical implications of the pluralism and externalism of the WMSH. I shall confine my attention to people who depart from normal moral agency. Questions of education obviously arise. Supposing that the WMSH is correct, are there things we can do to foster moral agency? This is particularly pressing when people are at risk or uncertain about what is right and what is wrong. Can externalist ideas about moral psychology be used to address risk and uncertainty? What are the implications for moral education? From here it is a short step to thinking of deeper deviations from normal moral agency. What about people who suffer from psychopathologies that impair moral agency? Do externalist ideas have therapeutic implications? I have no radical educational or therapeutic programs to offer. Indeed, I am skeptical of the prospects for satisfying moral education and therapy. Progress on these fronts will be made only in a piecemeal and tentative fashion.

Pluralism, amoralism, and practical application provide tools for assessing the overall width of moral minds. I have claimed that both narrow and wide mechanisms are at work in normal moral psychology. The obvious question to ask is "Which kind of mechanism is more pervasive?" Should we think of moral psychology as massively wide, or hardly wide at all? This is an empirical issue that cannot be resolved here. Indeed, given that empirical assessment of wide hypotheses is still an emerging practice, we should think that there is much work to be done before we can confidently form a picture of the relative width of the moral mind. However, we can survey the territory and try to bring the likelihoods into focus. Thinking about practical issues—the apparent opportunities for genuinely

wide contributions to education and therapy—is a preliminary way of assessing the relative width of the moral mind.

This chapter is even more speculative than the preceding ones. This is attributable, in part, to caution: it is notoriously difficult to move from the domain of theory to that of practice, so practical conclusions should be carefully drawn by those who develop theories. We philosophers have a pretty spotty track record here—we are not renowned for our practical acumen—so I will tread cautiously. The spirit of the discussion in this chapter stems also from recognition that the details of the phenomena discussed here are too complex to be adequately dealt with in one chapter. I have chosen to end on a speculative and programmatic note regarding these issues; fuller treatment must be sought elsewhere. All that I have to offer is some notes of caution about undue optimism about resolving hard educational and therapeutic challenges and some suggestions about places where efforts might profitably be directed.

6.1 Amoralism

Amoralism is, to put it simply and generally, insensitivity to morality. In what ways can we find ourselves either connected to or disconnected from morality? The amoralist is a stock player in philosophical approaches to moral psychology. Classically, the amoralist is someone who makes moral judgments but who is not moved by them. This figure is used to probe various intuitions about links between motivation and moral judgment. The so-called internalist holds, in some form, that moral judgment is necessarily motivating. The so-called externalist holds the opposite: moral judgment and motivation aren't necessarily linked, only contingently connected.[1] If the internalist is in some way correct, amoralism is conceptually impossible. If the externalist is correct, amoralism is conceptually possible. My present purpose is not to contribute to this debate. The pluralism and the embeddedness defended in connection with moral judgment in chapter 2 eliminate simple internalism and externalism as options for real people as envisioned by the WMSH. Insofar as internalism is committed to the claim that moral judgment necessarily motivates, the WMSH is best characterized as an externalist position.[2] Instead, my aim is to rethink the territory occupied by the amoralist: the bounds of morality. The figure of the amoralist straddles the boundary between normal moral agents and

those who are completely incapable of understanding morality. As someone who makes moral judgments and uses moral concepts, the amoralist understands moral considerations. Yet in being unmoved by these considerations, the amoralist is insensitive to morality in a way that normal moral agents are not. The amoralist makes judgments, as I do, but is as immune to these judgments as my cat is. This is interesting territory.

It is also more complex territory than this opening sketch of amoralism suggests. In what ways can people understand morality yet deviate from normal moral agency? Where do the psychological bounds of morality lie? The psychological conditions examined later in this chapter exhibit differing ways of being, or not being, a moral agent. Autistic people present various forms of deviation from normal moral agency; psychopaths present others. Are there yet other forms of deviation? Let us see.

6.2 A Look Around

Even if I have persuaded you that amoralism can function as a guide to moral psychology, you may still think it unnecessary to construct a taxonomy. After all, philosophers have discussed amoralism for years—surely the varieties of amoralism are well cataloged and understood. Unfortunately, this is not the case. Although philosophers have indeed discussed amoralism for years, not enough attention has been given to different versions of amoralism and to the implications of the variety of forms in which it comes. Here is a quick look at the messiness of this corner of philosophy.

The now-classic home of the amoralist is, as I have said, philosophical debate over internalism and externalism concerning moral judgment. Michael Smith (1994) discusses the amoralist in this way in his defense of internalism. However, broadening one's survey of philosophical discussions of the amoralist reveals a variety of conceptions of amoralism. In the internalism/externalism debate, the amoralist is characterized as making moral judgments. However, R.M. Hare (1981, 183) describes the amoralist as refraining from making moral judgments. Bernard Williams (1972, 1–2) characterizes the amoralist as challenging whether morality is rationally required of us. Such an amoralist might well both make moral judgments and act in accordance with them. The question here is whether the amoralist should act in accordance with these judgments. Richard Garner (1994)

defends a kind of amoralism that embraces a Mackie-style error theory about morality: moral judgments are assessable in terms of truth and falsity, and in fact they are all false. This sort of amoralist not only makes moral judgments, acts in accordance with them, and asks about the rational authority of moral norms, but also asserts a meta-ethical thesis about the status of these norms. These are four distinct notions of amoralism. This seems to imply that, instead of being the name of one kind of person, "the amoralist" might be an ambiguous name for several kinds of people. There seem to be many different sorts of amoralists among us, or at least up for our consideration.

One might object that, as presented, the contrast suggests a tension in philosophical treatment of the amoralist: seen one way, there may be no amoralists; seen another way, there may be many. The implication of this tension could be that philosophers are confused and in disarray when it comes to amoralism. One response to this objection is that there is reason to think that this tension is merely apparent: the debate between internalists and externalists about practical reasons concerns the extension of a particular kind of amoralism. In contrast, the disagreement found between apparently different discussions of amoralism concerns different ways of describing the amoralist. These are different issues, compatible with each other; the implied confusion is unwarranted.

This response is fair, so far as it goes. However, we should not blithely assume that inquiry into the extension of one notion of amoralism is insulated from other notions of amoralism. Take the internalism/externalism debate: this discussion has proceeded by linking the amoralist, described as making moral judgments without being motivated by them, to technical issues in meta-ethics and, in particular, moral psychology. However, there is no *a priori* reason to believe that other notions of amoralism have no implications for the same issues. The internalism/externalism debate has proceeded as if, in effect, there is only one relevant notion of amoralism.[3] Instead of being a safe assumption, the irrelevance of other forms of amoralism to the internalism/externalism debate is something that ought to be demonstrated. Generally speaking, we have reason to worry about judgments about the existence of one kind of amoralism that are made on grounds to which other kinds of amoralism seem to be relevant, but without consideration of these other kinds of amoralism. Likewise, we have reason to worry about judgments made on the basis of ideas about one

kind of amoralism about topics to which other kinds are amoralism are relevant.

Here is a reason for this state of affairs. Despite popping up with some regularity, the amoralist is rarely the primary topic of discussion. Instead, the amoralist gets considered as a secondary consideration in service of a different topic getting all the real attention, such as the nature of practical reason or the psychological foundations of morality. There is some reason to think that this method of proceeding has bred excessively facile handling of a potentially wide-ranging and important topic. For all its familiarity, whoever it picks out and whatever it means, "the amoralist" is the name of someone moral philosophers don't know very well.

6.3 A New Beginning

Let us begin as freshly as one can with a familiar topic. The most general way to characterize amoralism is as insensitivity to moral considerations. Specifying varieties of amoralism in further detail immediately bogs us down in such meta-ethical intricacies as those found in the internalist/externalist debate. This debate concerns a kind of person or a way of being: such amoralists can make moral judgments, but are not motivated by them. Since this sort of amoralism concerns how people (or other beings) are, let's call it *constitutional amoralism*. In contrast, the amoralists addressed by Williams and Garner need not be kinds of people. Williams's amoralist is someone who questions the rational necessity of morality. Instead of a way of being, this sort of amoralism is a way of thinking. Someone can entertain this sort of thought without changing the kind of person he or she is: it signifies a philosophical question, not a way of being. This kind of amoralism can be imagined as a rhetorical stance or strategy one adopts to make a point. Let's call this *stance amoralism*.

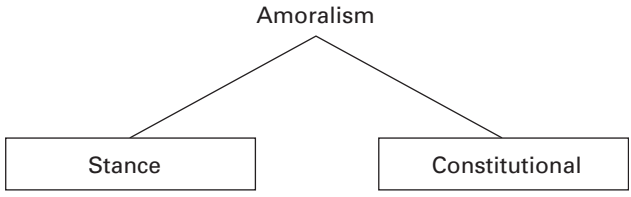


Figure 6.1

Although certain varieties of constitutional amoralism have received most of the recent attention, I will begin with stance amoralism.

6.4 Stance Amoralism

The kind of amoralist offered by Williams is a moral skeptic. In describing this amoralist stance, we need not require that the person who takes it is convinced that morality has no rational basis. All that is needed is the notion of an intellectual position or standpoint, the taking of which brings with it critical distance from morality. One takes the amoralist stance of the moral skeptic when one seriously inquires into the question of whether morality has a foundation in rationality, and in so doing refrains, for the purposes of a certain sort of inquiry, from making moral judgments of the sort that a person might usually be inclined to make.

The idea of stepping (perhaps temporarily) into the position of such a moral skeptic will be familiar to many academics, either from their own undergraduate days or from their experience with students who are willing to press the question of the rational basis of morality in conversation but who at other times act as convinced moralists. Less stereotypically, philosophers who, as professionals, wrestle with this question might well adopt this amoralist stance for a portion of their day in order to carry on a diligent sort of inquiry into the nature of morality.

When an amoralist stance is deliberately adopted, attention is directed toward morality in a way that interferes with our normal, unreflective connection with it. Such disruption can, presumably, have undesirable effects, but here we see the possibility that at least one form of amoralism can have benign or even good consequences. I will make more of this point when I turn to moral education.

Moral philosophy is not the only sort of inquiry the pursuit of which invites one to take the amoralist stance. Science is often represented, at least in public discussions, as amoral. The reason is that moral questions are not among those that are definitive of scientific domains (putting aside moral psychology, of course).[4] Take chemistry as an example. Arguably, when a chemist enters a lab to inquire into the chemical mechanisms of some phenomenon, moral questions about that phenomenon become a secondary focus at best. It makes no difference to the nature of, e.g., the chemical processes that occur when oil from a ruptured tanker gets into

the feathers of a seabird whether the oil company is morally responsible for the spill or not.[5] Nor does it make any difference to the description of these processes. Moral questions are not among the structuring concerns of chemistry as a domain of inquiry. Consequently, chemistry is plausibly seen as amoral in content, and a person who works as a chemist is plausibly seen as taking an amoral stance when he or she actively pursues this activity. The same goes for many other sciences.[6]

However, there is an important difference between the sciences and the moral skepticism involved in pursuit of certain sorts of philosophical concern. Arguably, the relation between science and morality is more complex than the sketch I have provided here. It is, to some degree at least, a matter of substantial debate whether the sciences must be amoral. In fact, it is not uncommon for commentators to note that value judgments, including clearly moral ones, do seem to be a part of much scientific activity. Howard Slaate has usefully surveyed some of the relations between science and ethics. Many of the connections stem from the practical side of science (Slaate 1981, 156–163). Others result from the fact that scientific inquiry takes place within complex institutional settings that require decisions about the allocation of resources. The notion that science is amoral is most persuasive for the so-called pure sciences—that is, the sciences whose only purpose is to produce knowledge, which is not necessarily to be applied. However, even the results of pure science can be applicable, and pure science clearly is subject to the same institutional issues as more directly practical science. Since there seems to be a substantial question here, and since there is reason to think that moral questions are interwoven with the practice of science in complex ways, I think it is fair to characterize the stance taken when investigating scientific questions as *contingently* amoralist. In contrast, the moral philosopher who inquires into the question of whether morality is rationally required of us cannot avoid taking an amoralist stance. It is a necessary structuring feature of this sort of question that one "bracket" morality. Academic moral skepticism, therefore, is best seen as *necessarily* amoralist.

What does the psychology of stance amoralism indicate about normal moral agency? It is obvious that reasoning is a precondition of adopting these sorts of stances. This presents us with substantial and methodological lessons. Moral reasoning is constitutive of normal moral agency, yet

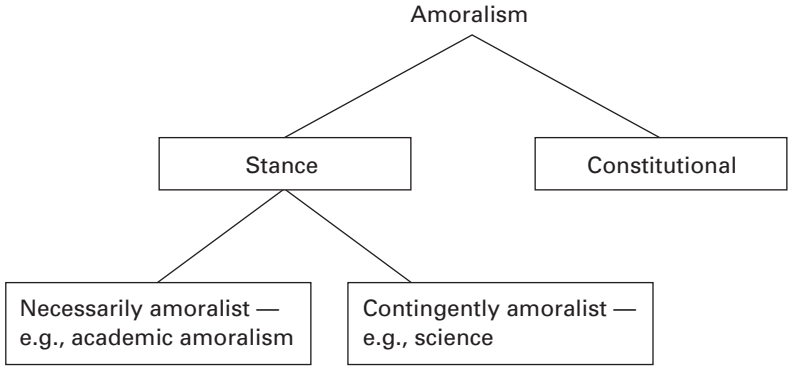


Figure 6.2

reasoning is also needed, *prima facie*, to achieve one sort of insensitivity to moral demands. Indeed, it might very well be reasoning about moral issues that prompts one to ask the question characteristic of Williams's amoralist. The substantial lesson here is that we should not assume that there are psychological processes that guarantee normal moral agency. One and the same process might be capable of fostering and eroding normal moral sensitivity. The methodological lesson is that, for any psychological process identified as necessary for some variety of amoralism, there are two questions to ask. First, we should ask how it delivers the lack of sensitivity to morality. Does it interfere with some distinct psychological process or processes needed for moral agency, or does it operate in some other way? Second, we should ask whether the process in question is a part of our normal moral psychology. We can imagine amoralists with these capacities, but can we imagine competent moral agents who lack them? If not, then the sort of amoralism in question will have been of indirect aid in pinpointing a particular feature of normal moral psychology and, hence, must be addressed by any complete and adequate theory of moral psychology.

We already know that reasoning is a part of normal moral agency. What about the first question: How does reasoning deliver stance amoralism? Here are two suggestions. First, imagination seems to be needed. The reason is that moral agents must be able to think counterfactually in order to attain stance amoralism. That is, they have to be able to think about

what it would be like to be different from the way they are now. Imagination delivers the idea of the stance to be occupied. Actually occupying the stance requires more than imagination. The reason for this is that we seem to have fairly automatic processes of judgment, action, and feeling that constitute normal moral agency. If one is to occupy an amoralist stance, these normal processes must be inhibited somehow. Perhaps it is impossible to turn these off, but their activity can be sufficiently suppressed to make it ineffectual. Perhaps they can be temporarily rendered inoperative without being completely dismantled. The truth here awaits empirical assessment, so I shall not advance hypotheses with much detail or confidence. At this point, it is safe to think that adopting an amoralist stance requires self-directed processes that inhibit the normal features of moral agency in a controlled manner. These processes could be plural and domain specific—for example, we might require multiple mechanisms to suppress the various ways that moral judgment and feeling normally occur. Alternatively, there might be one higher-order process of self-control that accomplishes all of the necessary suppression. I will not choose between these options.

Imagination and self-directed, controlled inhibition are putatively necessary both to posit and to occupy an amoralist stance. Is either of these also constitutive of normal moral agency? I am inclined to think that both of these capacities are included in our mature moral psychology. Imagination is needed for understanding the lives of those who live in much different ways than oneself. This goes for both human and nonhuman recipients of one's thought and action. Imagination is also needed for thinking abstractly about moral issues and values. Thinking about imaginary cases is a common feature of university moral education. Self-directed, controlled inhibiting processes will be necessary to the degree that one has tendencies that either undermine normal moral agency or lead one to do wrong despite being a competent moral agent. Those who are pessimistic about the depths of human goodness are likely to think that such processes are of central importance to normal moral agency. Optimists about human nature will be more inclined to downplay them. To the extent that we are characterized both by good and bad tendencies—i.e., to the extent that human nature is fragmented in complex and normatively inconsistent ways—we will also need self-directed controlling processes in multiple forms in order to pursue the good.[7]

6.5 Constitutional Amoralism

Constitutional amoralism is the sort of amoralism that, at least in part, characterizes one as the sort of being one is. The most obvious form of this occurs with people who are different, psychologically, from mature, reasons-sensitive adults. The psychology of such paradigmatic figures includes the capacity to be sensitive to moral reasons. In contrast, other psychological configurations preclude such sensitivity or put serious obstacles in its way. In general, let's call this *psychological amoralism*.

The most obvious kind of psychological amoralism involves *abnormal* cognitive abilities; that is, some sort of impairment interferes with otherwise normal moral cognition. This, in general, is the case with psychopathy, at least in its most familiar form. The psychopath seems to be capable of recognizing moral reasons, but is psychologically constituted so as to be insensitive to them. (The psychology of psychopathy will receive extended attention later in this chapter.)

It would be a mistake to limit psychological amoralism to cases of abnormal psychology resulting in psychopathy. Besides impaired psychology, being psychologically immature is another way of being that precludes sensitivity to moral reasons, but psychological immaturity is a perfectly normal stage of development for all humans. Infants are amoral: they are not sensitive to moral reasons, and we do not treat them as wholly morally

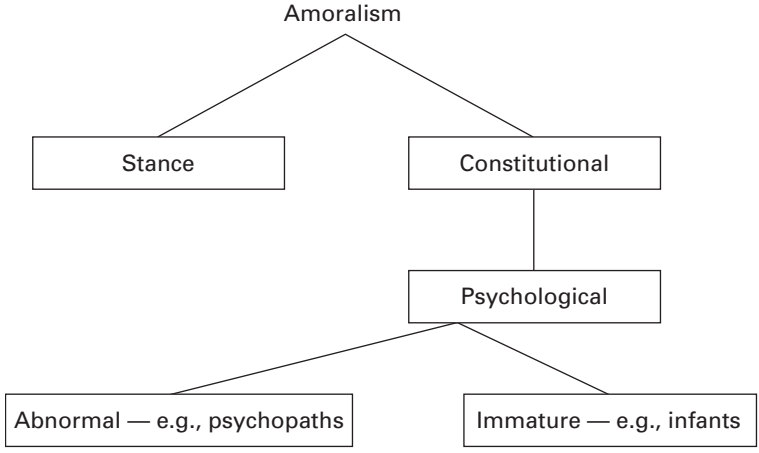


Figure 6.3

responsible for their behavior. Such amoralism is due to psychology, but infants are not psychologically abnormal or impaired. They simply aren't ready to be full participants in moral reasoning, attributions of moral responsibility, and so on. This is quite different from psychopathy.

Infants are amoral in two ways. First, they are not responsive to moral reasons. Second, their conduct is not subject to moral assessment. Infants do neither right nor wrong. Let's call the first sort of amoralism *receptive* and the second sort *productive*. Receptive amoralism does not imply productive amoralism. They come apart in the cases of at least some psychopaths. Such people are not responsive to moral reasons, but their conduct is, at least at first glance, still properly assessable in moral terms. Without delving into the matter very deeply at this point, it seems to me to be likely that productive amoralism does imply receptive amoralism. Infants provide the clearest case[8] of productive amoralism, and they are also receptively amoral. Some people with abnormal psychologies are also productively

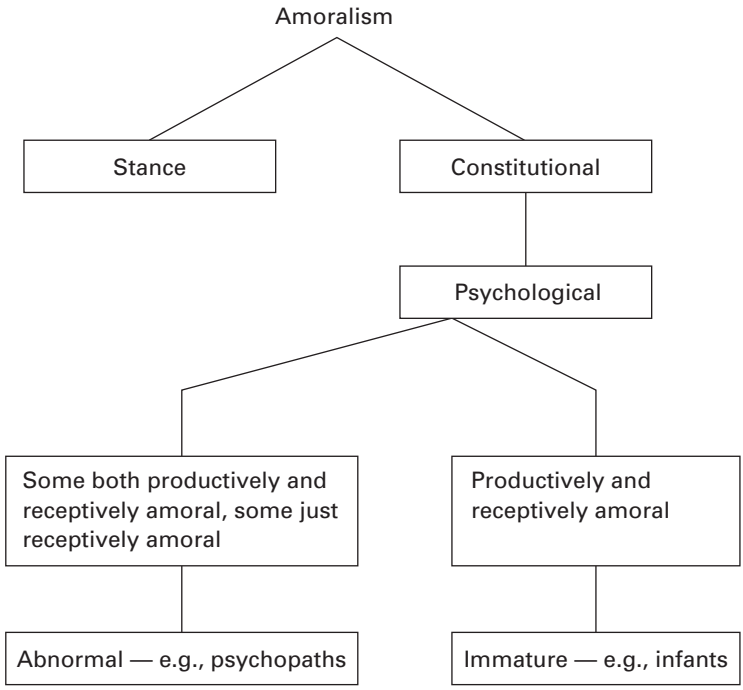


Figure 6.4

amoral, and I expect that they will be receptively amoral also. Strictly speaking, since this is a matter of the nature of cognitive, affective, and action-producing capacities, whether productive amoralism always implies receptive amoralism is a matter that must be investigated empirically.

The distinction between receptive and productive amoralism helps us to navigate the issues that arise with stance amoralism. Let's return to chemistry. Insofar as moral concepts and questions do not structure the domain of chemistry, a chemist's judgments *qua* chemist are not responsive to moral reasons. Such a stance is receptively amoral. But the activities performed in the name of such a stance are plausibly taken to be subject to moral assessment. Moral problems arise in connection with the amoralism of scientific stances when practitioners implicitly or explicitly behave as if their stances are *productively* amoral. This is deeply implausible. The inclination to think so might derive from a failure to differentiate varieties of amoralism or from the implicit assumption that receptive amoralism implies productive amoralism.[9]

Does psychological amoralism add anything to our account of what a complete theory of moral psychology should include? Specific impairments—psychopathy and autism—will be dealt with later in this chapter.

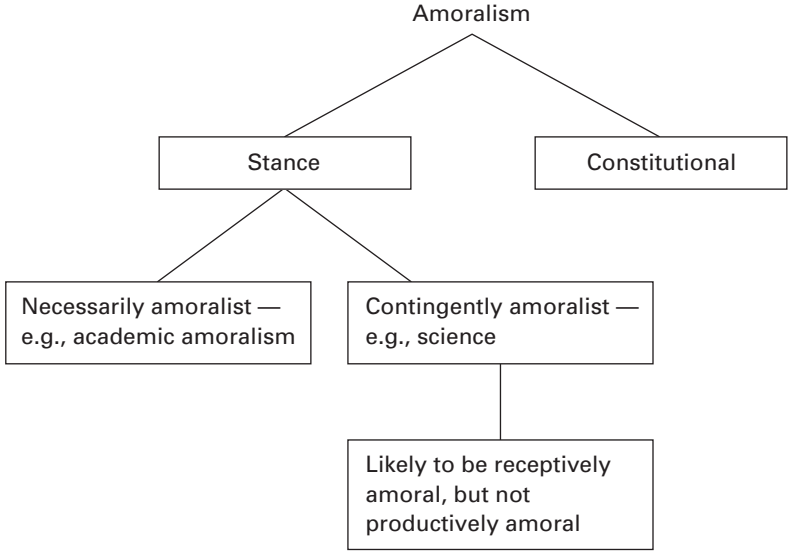


Figure 6.5

Immaturity by itself offers nothing new, because it applies to moral psychology as a whole: *whatever* characteristics constitute normal moral agency, infants will lack them, at least in their mature forms. Developmental psychology will continue to have much to contribute to our understanding of the mechanics of the moral mind, but my suspicion is that productive study into psychological development must generally follow accounts of mature psychology, not lead them. Hence, although psychological amoralism is an important branch of the present taxonomy, it offers little for us to add to our map of normal moral psychology.

Psychological amoralists are amoral as a result of their psychological capacities. However, psychological functioning is not all that contributes to the constitution of ways of being. Commitments and convictions are also components of ways of being. This is related to the taking of a stance, yet distinct from it. A stance is something one can adopt or leave voluntarily.[10] Convictions and commitments are typically more resilient. They can be altered, but not easily; to say that one can give them up voluntarily is misleadingly facile, as it might takes years of painful effort to shed a commitment or conviction. For the purposes of contrast, let's say that psychological amoralists are amoral because of the way their cognitive systems work. In contrast, people who are amoralists as a result of commitments or convictions are amoral as a result of adherence to propositions. Such adherence constitutes the kind of character such people have. I call this *propositional amoralism*.

There is some reason to think that the varieties of amoralism that figure the most in current meta-ethics, at least under the name "amoralist," are

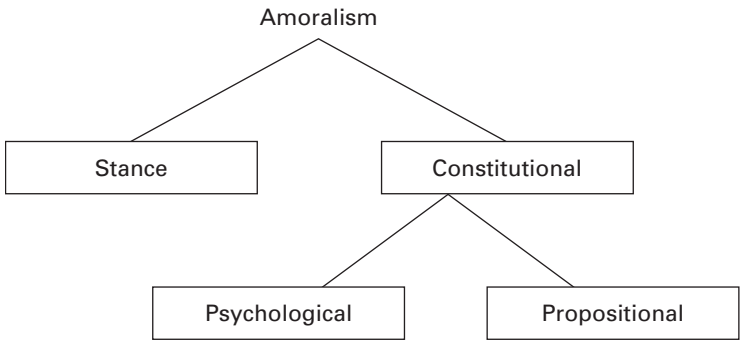


Figure 6.6

versions of propositional amoralism. Generally, such versions of amoralism concern whether one should be a certain way or not, or whether one should believe or act in accordance with certain propositions or not. Consequently, they are the sorts of things about which one can ask whether the possibilities they represent are coherent or sensible, or whether choosing amoralism of one of these forms is rational. These are the sorts of questions with which analytic meta-ethicists are most comfortable. However, given that we have already seen two other varieties of amoralism, I hope that some doubt has already been raised about such preoccupation.

Based on the kinds of amoralism offered by present-day meta-ethicists, we can distinguish two broad families of propositional amoralism. As I have already noted, Richard Garner defends a version of Mackie's error theory about morality under the name of amoralism. Such a character thinks that moral discourse is, strictly speaking, false: there are no moral facts of any kind, no true moral sentences, no binding moral prescriptions. Such is the position of a thoroughgoing moral skeptic. Given that all aspects of morality are touched by such a position, I shall call this *complete propositional amoralism*. In contrast, the sort of amoralist discussed in connection with internalism and externalism about moral reasons is represented as making moral judgments. Such a person can countenance moral facts. However, an amoralist of this sort is unmoved by moral considerations. Given that the difference between an amoralist of this sort and a typical moralist is a difference against a background of (at least potential) agreement about morality, I shall call this *partial propositional amoralism*.

Because literature on internalism and externalism about moral reasons is lengthy and complex, I could list many subtly distinct versions of partial propositional amoralism. I shall confine my attention to two.

Generally, the partial propositional amoralist is not moved by moral considerations. There are (at least) two forms that such practical insensitivity could take. First, such an amoralist might recognize moral facts but deny that they provide reasons for action. I call this *justificatory partial propositional amoralism*. Second, such an amoralist could recognize moral facts and admit that they provide reasons for action, but fail to be moved by these reasons. I call this *motivational partial propositional amoralism*. Fiction provides examples of such figures. The principal characters in the television series *Seinfeld* seem to me to be partial propositional amoralists

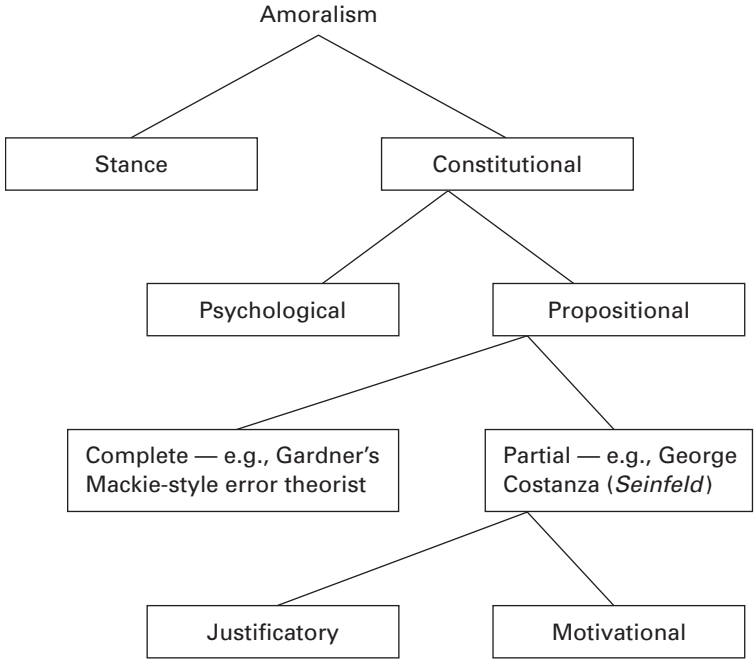


Figure 6.7

(some of the time): they seem to recognize moral facts, but either they see them as providing no reason to act or they are not moved by the moral reasons they recognize.

The distinction between justificatory and motivational amoralism gives us a tool with which we might discern more finely graded versions of psychopathological amoralism. The psychopath is insensitive to moral considerations. This could be in a justificatory sense: the psychopath does not see moral considerations as providing reasons for action. Alternatively, it could be in a motivational sense: the psychopath recognizes moral reasons, but is not moved by them. Insofar as these insensitivities result from psychological impairment, it is tempting to think that justificatory psychopathological amoralism, if actual, might result from cognitive impairments, whereas motivational psychopathological amoralism, if actual, might more plausibly be attributable to affective problems. However, owing to my pluralist leanings, I suspect that the relevant mechanisms are more numerous than this, and the details more

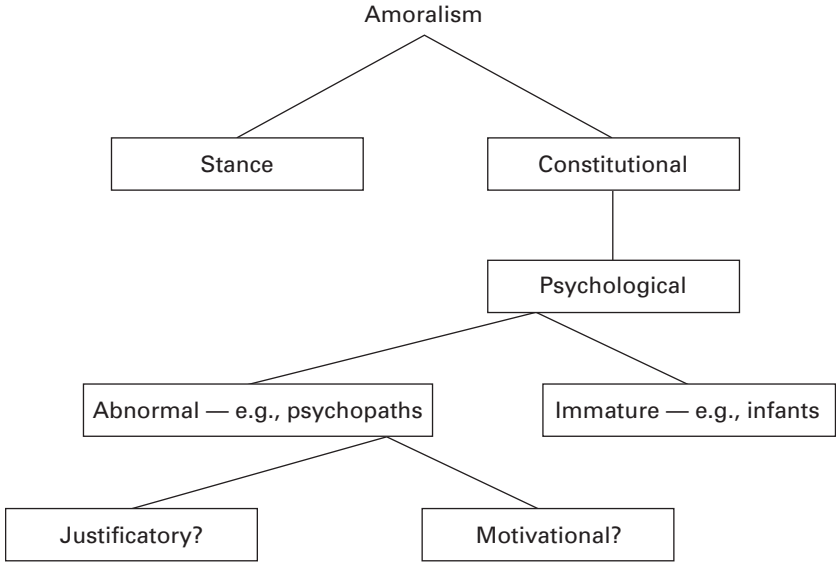


Figure 6.8

messy. Both the actuality of these sorts of amoralism and the mechanisms that bring them about are to be decided by empirical psychology, not by philosophy.

What does propositional amoralism tell us about normal moral psychology? I shall postpone psychopathy until section 6.13. The general motivational issues fall into the category of the production of action, the topic of chapter 5. They are also related to moral judgment, to moral reasoning, and to the attribution of responsibility. The distinctive feature of propositional amoralism is the attention it draws to committing oneself to a principle or position. Garner's amoralist is distanced from morality by virtue of a metaethical commitment. Commitments that are more constrained in scope might bring about more circumscribed insensitivities to moral demands. To my mind, the most striking thing about this form of amoralism is that it is brought about by a phenomenon that is regularly taken to be a feature of normal moral agency. Both the amoralist and the moral saint can, apparently, have strong commitments. Moreover, it is these very commitments that make them the propositional amoralist and the saint, respectively. Insofar as the psychology of committing oneself to a position or principle is not obviously accounted for by discussions of

moral judgment, reasoning, and emotion, it promises to offer something to be added to any adequate moral psychology.

In just what psychological processes does committing oneself to a principle or position consist? The speculative answer that can be offered at this point mobilizes ideas similar to those suggested in my discussion of stance amoralism. This is not surprising, in view of the similarities between taking a stance and designing one's life around certain ideas. Imagination is, I suspect, more important to taking a stance than to committing oneself to a principle. Insofar as one performs abstract moral reasoning about imaginary cases on the basis of one's commitment, imagination will be required. But it seems to me that one can commit oneself rather blindly, without planning, forethought, or imagination. If there is a need for imagination in committing oneself, it is for assessing one's success at living up to one's commitments. Suppose I am trying to design my life around a principle of self-reliance. To be successful, I must keep my behavior, and perhaps my thinking, in line with my principle. Somehow my behavior must track my commitment. If it does not do this, any success I have in executing my commitment will be attributable to luck; the likely outcome will be that I will fail to live up to my commitment. One way of tracking my commitment is by conscious assessment of my behavior in light of my commitment. This seems to require imagination, as I have to be able to think about counter-factual situations. I must be able to imagine the sort of behavior consistent with self-reliance. With this information in hand, I can measure my own behavior against what I have imagined. However, I see no *a priori* reason to think that the tracking of a commitment must be done through conscious processes. Presumably people can develop habits of thought and action, and these can effectively instantiate principles in the absence of conscious imaginative assessment of one's own behavior. I take it that such unconscious patterns of thought and behavior are characteristic of what is sometimes referred to as a moral sensibility.

All of this suggests that processes of self-regulation are necessary for commitments, even if imagination is not. As with taking a stance, inhibition will be important to self-regulation. However, it is likely that the self-regulation characteristic of commitment involves more constructive self-directed processes than stance-taking does. A commitment to being charitable, for example, is a commitment not only to avoid certain ways of acting but also to act in other-directed ways. It might also be a commit-

ment to condition one's outlook—the way one sees and thinks about the world—so as to avoid certain ways of thought and to pursue others. Such self-regulation probably involves both emotional and reasoning processes. Some of these might be very domain specific; others might be quite general in application.

A third sort of constitutional amoralism is akin to psychological amoralism, since it results from the way mechanisms of decision making are arranged. It is similar to propositional amoralism in that it is better seen as a matter of character than as a matter of constitution. Its most important feature that it is not a form of amoralism that is a possibility for individual humans. Instead, it is the sort of amoralism that can characterize corporations, governments, and other semi-formal collections of humans. I call it *institutional amoralism*.

Generally, institutional amoralism is a form of receptive amoralism: it occurs when institutional decision making takes place in such a way that moral reasons are not included. Clearly, moral problems can arise here that are very similar to those that can arise with scientific stance amoralism: the risk is that receptive institutional amoralism is taken to imply productive institutional amoralism. We succumb to this risk if we think that institutional activities are free from moral assessment. This is implausible: we regularly criticize government activities in moral terms without this seeming deeply mistaken, and the literature on corporate responsibility has developed in such a way as to make principled theoretical room for the moral assessment of corporate activity.

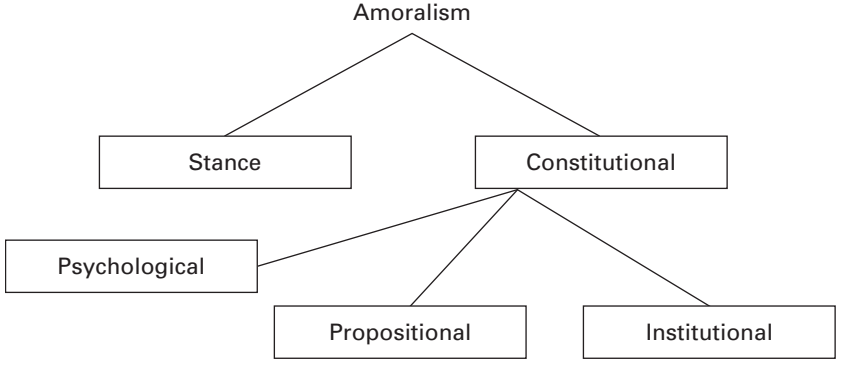


Figure 6.9

It seems to me that there are two broad ways in which institutional amoralism can come about. First, it can be the intentional result of the structuring of the mechanisms of decision making for the institution in question. Institutions can, in principle, structure themselves to leave moral considerations out of institutional decision-making processes. Let's call this *planned institutional amoralism*. On the other hand, as institutions evolve, and especially as decision making is spread among people or functional units of a given institution, it might just come to pass, without planning, that moral considerations are left out of institutional decision making. I call this *contingent institutional amoralism*.

The psychology of planned institutional amoralism probably involves two things already addressed: the coordination of behavior with others and the psychological mechanisms needed for committing oneself to an idea or position. Social context provides a natural home for widely realized versions of these processes. Contingent institutional amoralism is different. It need not have any implications for individual moral psychology at all. Since such amoralism is not planned, it could be the case that an institution evolves by chance to ignore moral considerations even when

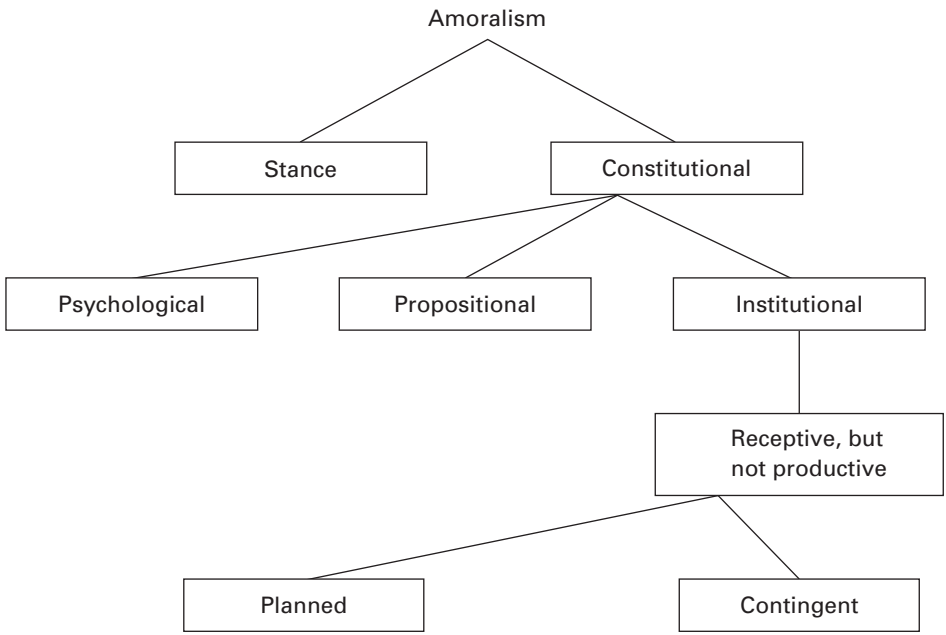


Figure 6.10

all the people who execute the activities of the institution are normal moral agents. However, it might be that certain individual tendencies make contingent institutional amoralism more likely. Social processes effecting conformity of thought and judgment in ways that ignore or downplay moral considerations are probably important for this phenomenon. The same goes for the social processes that lead to distribution of responsibility and joint shaping of behavior, as discussed in chapters 4 and 5. Although empirical examination of such processes is still needed, I suspect that nothing new is added to our list of components of moral psychology by institutional amoralism.

6.6 Immorality and Anomalous Amoralism

Although I have mentioned in passing the moral assessment of the actions of some sorts of amoralist, the matter of immorality has received no direct attention. Some might find this odd owing to the use of "amoral" as a synonym for "immoral."[11] Of the sorts of amoralism so far cataloged, I do not think that any automatically imply immorality. However, some might make immoral acts more likely. For present purposes, the issue of immorality points toward two new kinds of amoralism, both trickier than those so far examined. Since they fit a bit uneasily into the scheme so far developed, I shall call them *anomalous*.

In *Paradise Lost*, Milton represents Satan as opting out of traditional morality: "Evil be thou my good." This makes Satan an amoralist. Since it is a matter of character and moral conviction, Satan's amoralism is a variant of the propositional family. However, instead of being absolutely insensitive to morality, Satan is responsive to an inverted version of it. Instead of seeing, for example, the production of pain as a reason to avoid doing a certain action, for Satan it will be a reason to perform it. This is paradigmatic immorality, but, as it involves a certain sort of rejection of morality, I think it counts as a version of amoralism as well. For taxonomical purposes, let's call this *evaluative anomalous propositional amoralism*.

The second kind of anomalous amoralism is also a version of propositional amorality. Instead of choosing evil, such a person chooses lack of value, perhaps even complete nothingness. The protagonist of the Velvet Underground song "Heroin" decides that he is going to nullify himself, not with death, but by opting out of interpersonal activity, and of most self-directed behavior, by chemical means. In nullifying his life, he attempts

to commit himself to a void, so to speak. He tries to commit himself to nothingness, including the idea that anything matters. Since the choice is to renounce the realm of value altogether, it implies insensitivity to moral reasons, and hence counts as a form of amoralism. I call this *null anomalous propositional amoralism*.

The null and evaluative versions of anomalous amoralism are tricky because I am not sure that we can actually fulfill them. On one hand, I suspect that, in trying to become such characters, either we would merely fail or we would die in the process (at our own hands or at the hands of others). On the other hand, I am not convinced that these are within the range of live psychological options for normal humans. Unlike the previous forms of amoralism, these versions call for empirical assessment in order to assess their very existence. Only after such confirmation would it be worthwhile to assess their psychological roots and their implications for normal moral agency.

This ends the tour of amoralism. Figure 6.12 presents the entire taxonomy.

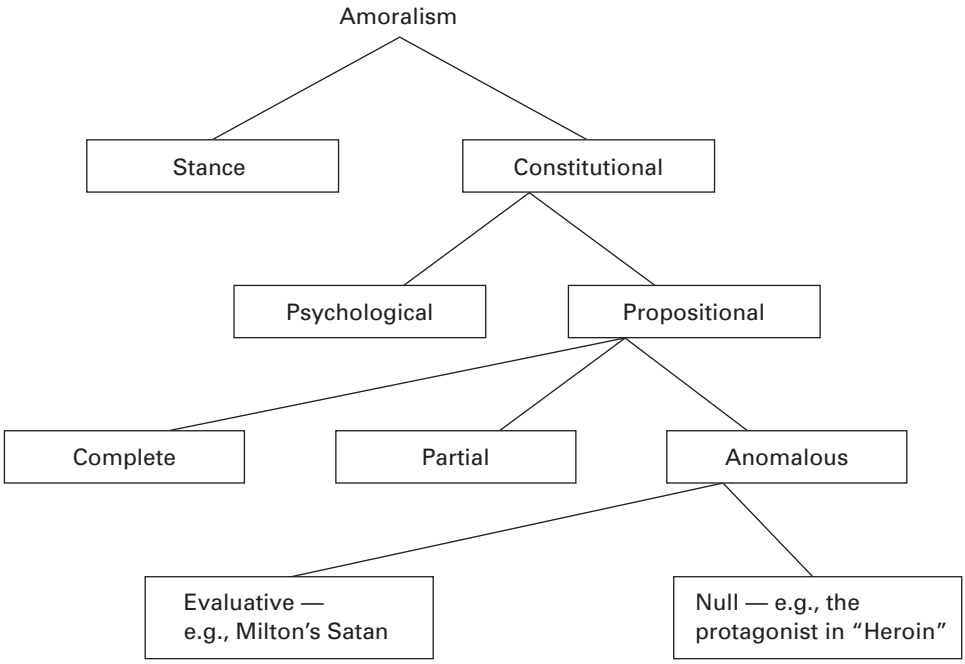


Figure 6.11

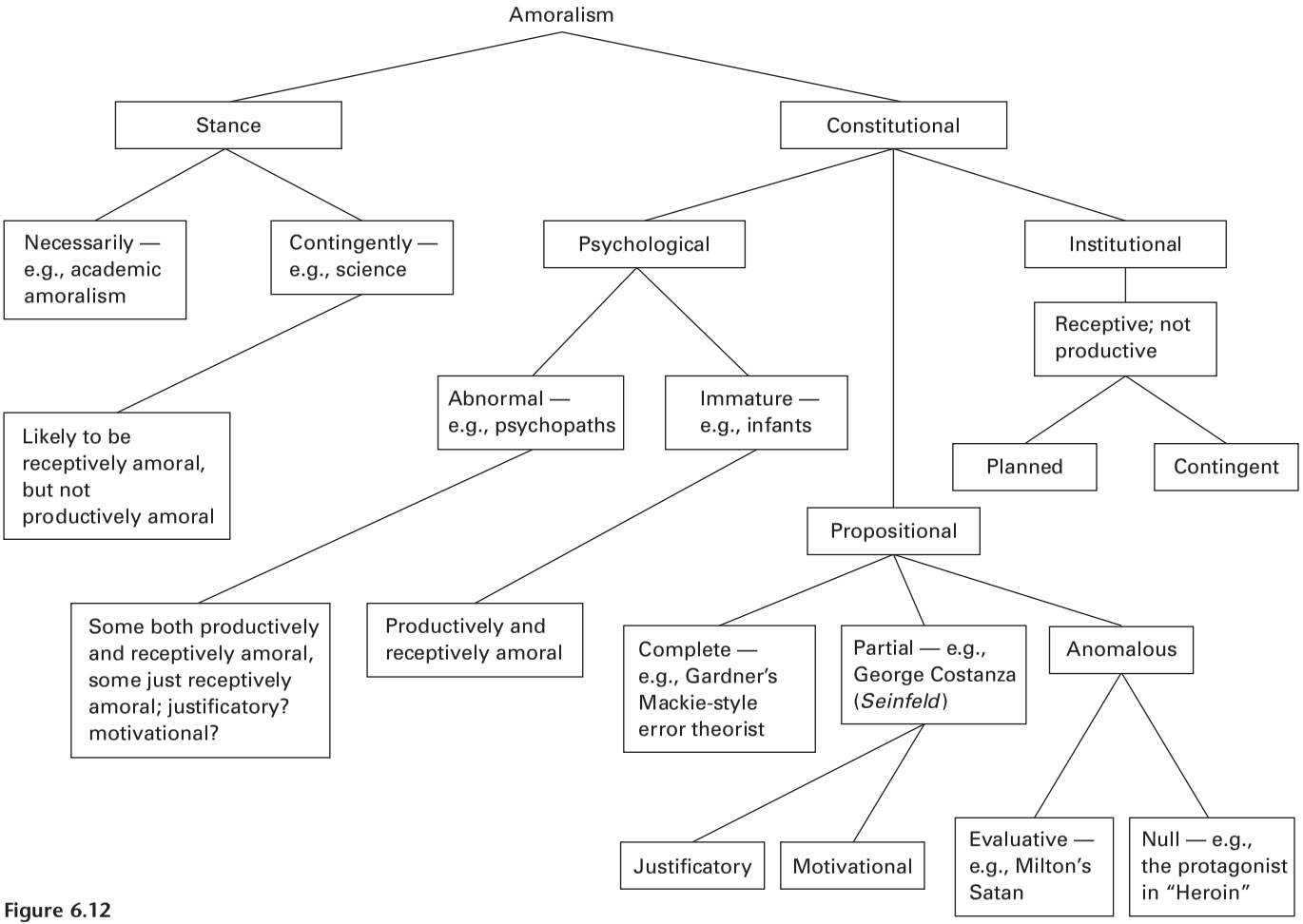

Amoralism
Stance
Constitutional
Necessarily — e.g., academic amoralism
Contingently — e.g., science
Psychological
Institutional
Likely to be receptively amoral, but not productively amoral
Abnormal — e.g., psychopaths
Immature — e.g., infants
Receptive; not productive
Planned
Contingent
Propositional
Some both productively and receptively amoral, some just receptively amoral; justificatory? motivational?
Productively and receptively amoral
Complete — e.g., Gardner's Mackie-style error theorist
Partial — e.g., George Costanza (*Seinfeld*)
Anomalous
Justificatory
Motivational
Evaluative — e.g., Milton's Satan
Null — e.g., the protagonist in "Heroin"

Figure 6.12

6.7 An Explanation of the Heterogeneity of Moral Psychology

It is not hard to think of ways in which my catalog of varieties of amoralism is likely to be expanded. Both conceptual refinement and empirical addition are likely to yield varieties of amoralism that are not explicitly represented in my introductory taxonomy. A particularly promising idea that has not been pursued here is moral experience. Perhaps there are phenomenological aspects of moral experience that can be used for adding one or more varieties of amoralism. This probably would develop the remarks made about moral sensibilities, with a focus on the question of whether varieties of normal moral experience have distinctive phenomenological features. In a vaguely related vein, perhaps additional branches will be categorized in terms of emotional aspects of moral thought and experience. For instance, if I am correct that evaluative and null anomalous propositional amoralism are psychologically impossible for us, the explanation for this might be that our affective propensities make them impossible. Regardless of future additions, what is presented here will suffice for my present purposes.

It should be clear that we stand in various psychological relations to morality. This opens up the possibility, charted in this section, that we can deviate from normal moral agency in multiple and complex ways. My hypothesis has been that, although some of our connections to morality are individualistically realized, some of the psychological pillars of morality are wide, comprising both bodily bound and environmental cognitive resources. Although throughout this book I have emphasized conformity and mind-reading capacities as central to a wide view of moral psychology, I would wager that other sorts of external resources and psychological capacities for making use of them could be revealed through examination of some overlooked forms of amoralism. Stance amoralism strikes me as particularly ripe for externalist exploration, but this will have to be done elsewhere.

It should be no surprise that normal moral agency is psychologically heterogeneous in the sense of being constituted by moral judgment, moral reasoning, emotion, and mechanisms for producing action. More of a surprise is the diversity of mechanisms that seem to be at work realizing each these components of moral agency. What might explain this plurality? I cannot offer a complete answer, of course. I doubt that there is an answer that offers a rationale for all aspects of this pluralism, as

some degree of it probably is due to the contingencies of our natural history. This is just how it worked out through human evolution. But I think a partial answer can be offered. In a word, the explanation for this pluralism is externalism. In two words, the explanation is environmental reliance. The world in which we operate offers many resources ripe for inclusion in cognitive processes. Not all such incorporation will be accomplished through use of the same sort of processes. Not every use of an environmental resource will be efficient or elegant; thus, very similar worldly features might be used by distinct parts of our mental make-up. As with individualistic views of the mind, we should expect some degree of redundancy, maybe even a considerable degree, in the ways in which we use the world. In short, the moral mind is composed of heterogeneous processes because the world in which it operates offers diverse resources.

The individualist about moral psychology can offer a similar explanation of the heterogeneity of the psychology of moral agency, but not exactly the same explanation. Besides taking account of contingency and redundancy, the individualist can explain the pluralism of moral psychology in terms to the psychological jobs to be accomplished. This goes for both the higher and the lower levels at which pluralism has been described in this chapter. For instance, consider moral judgment. Suppose that an individualist posits a process by which agents make moral judgments by analyzing certain features of actions, in the spirit of Hauser's account. Combined, the fact that we make moral judgments in this way and the fact that we live together in complex ways point to a second psychological job that might need to be done: competition and the benefits of comfortable social interaction mean that it might be important to conform at least some of our moral judgments to those made by our conspecifics. This can be accomplished in various individualistic ways. One possibility is that the mechanism already posited can be subverted, so that sometimes instead of using features of actions it delivers judgments on the basis of information about the views of others. Alternatively, a second mechanism of moral judgment can be posited: the first mechanism judges on the basis of action features, the second on the basis of the views of others. As psychological jobs are added, hypothetical mechanisms can be multiplied.

The externalist can, in principle, posit all the mechanisms offered by the individualist. But the externalist can also account for pluralism in two

other ways. Let's continue with moral judgment, but let us also note that the general pattern applies elsewhere too. Besides all the individualistic hypotheses, the externalist can surmise that agents make moral judgments by using some of the cognitive resources offered by other people, including their moral judgments. Such environmental exploitation can be accomplished in various ways, so this avenue of explanation need not add only one more mechanism to the collection offered by the individualist. Second, the externalist can deploy radically different sorts of resources from those available to the individualist. For example, if I am correct about the importance of mind reading and processes of effecting conformity of thought and behavior to the likelihood of the truth of wide hypotheses, then at a simple limit moral judgments might be produced *solely* by these processes. Whereas the individualist must offer a judgment-producing mechanism in addition to these, the externalist need not.

Two things are worth noting about these patterns of explanation. First, the ability to explain pluralism is not necessarily much of a theoretical virtue. The extent to which this is a desirable theory of an account of moral psychology will depend on the data. My impression is that, as data come in, they are pointing to increased psychological pluralism. To the extent that this is correct, the explanation of such pluralism becomes more interesting and important. Second, and relatedly, this issue provides a perspective from which to reflect on the relative width and narrowness of moral psychology. Externalism accommodates more psychological heterogeneity than individualism does. Hence, the greater the extent to which the data provide grounds for positing multiple mechanisms for each component of our moral psychology, the more *prima facie* support there is for externalism about moral agency. This is indirect support, since the crucial issue is not the number of mechanisms but their constitutive dependence on or independence from worldly resources. Still, let me suggest that pluralism about the mechanisms of moral judgment, moral reasoning, and so on provides reason to take externalism about these capacities seriously.

6.8 Assessing the Implications of the Externalism and the Pluralism of the WMSH

Both the externalism and the pluralism of the Wide Moral Systems Hypothesis result from the psychological sensitivity of agents to features of their

environment. So far, I have been concerned with the theoretical implications of this environmental sensitivity. However, it is reasonable to expect that it should have practical implications as well. In the remainder of this chapter, I will examine environmental sensitivity in connection with adult moral education, with autism, and with psychopathy. All three raise more pointed and, in the cases of autism and psychopathy, painful issues about insensitivity to moral demands than the more abstract considerations presented in the taxonomy of amoralism. Before we turn to these tricky practical issues, environmental sensitivity itself deserves attention.

6.9 Varieties and Methods of Environmental Sensitivity

Our environment is, by any standard, rich in diverse types of information. How do we make use of this information? By what means are we sensitive to it? A simple distinction is fundamental both to understanding these issues and to seeing important possibilities for education and therapy. First, there are well-known domain-specific modes of environmental sensitivity. Our senses provide examples with which even very young school children are familiar. Our eyes—or, more properly, our visual systems—allow us to detect and make use of things that can be seen; our ears put us in touch with information that can be heard. Subtler examples can be added to the elementary school child's inventory. Perhaps emotions are perceptual capacities by which we track things that matter to us. Our mind-reading capacities allow us to detect the thoughts of others. We have an impressive array of specialized tools for detecting and making use of particular sorts of information supplied by our environment. In contrast to these first-order, domain-specific varieties of environmental sensitivity, we also have second-order capacities that appear to be domain-neutral. 'Attention' is the familiar word we give to our capacity to focus our minds on something in particular.[12] The focusing of the mind seems to be something I can do either with sights or with sounds or with other minds, the same capacity at work across the range of first-order capacities. Whether this is really the case must be left to more detailed empirical study of attention. For present purposes, all we need is the distinction between domain-specific and domain-neutral varieties of environmental sensitivity. Suppose that these are our raw materials. What can be done to foster and improve performance for a particular kind of environmental sensitivity? Broadly speaking,

we can distinguish three sorts of interventions we might make. First, we can *facilitate* sensitivity to some sort of environmental information. This probably takes a variety of forms, foremost among them putting in place the means of sensitivity and removing obstacles to its functioning. Second, we can *supplement* environmental sensitivity in the hope of making it work better. Finally, we can *replace* a form of environmental sensitivity with another form that may perform better or may make up for some sort of deficit. In principle, these three sorts of intervention apply equally to domain-specific and domain-neutral forms of environmental sensitivity. However, in practice our options are more constrained. First, the development of domain-neutral means of education and therapy will probably be the most initially attractive option. Second, specific sorts of domain-specific means of education and therapy will emerge as more promising as the contingent details of the workings of our minds emerge from the hard work of empirical research. Third, possibilities for environmental modification will emerge as ways of enhancing our environmental sensitivity without changing individuals.

It is one thing to identify such general possibilities for education and therapy. It is quite another to implement them successfully. I will look at some domain-neutral and domain-specific possibilities when I turn to autism and psychopathology. First, to illuminate some of the pitfalls here and hence to temper our expectations for education and therapy from a philosophical theory about our moral psychology, let's examine attention and adult moral education.

6.10 Educational Implications of the WMSH: The Vagaries of Moral Reasoning 101

Perhaps you or someone you know earns a living, as I do, by teaching moral philosophy in a university. Perhaps you took such a course as a university student. It is not uncommon for people who have taken such courses to wonder what their point is. Certainly they have internal connections to the programs in which they are located, but so do lots of other courses. These philosophy courses seem different from such other courses because of their subject matter. They are more intimately linked to the everyday concerns of ordinary people than courses in, e.g., English literature, biology, math, or other areas of philosophy. And yet my students

seem to emerge from my courses largely unscathed by my teaching. So far as I can tell, they have neither an academic interest in moral philosophy, nor a new appreciation for the nature and the complexity of values and familiar moral issues, nor a new practical outlook on their lives. What's the point? Supposing that I could answer this question, would a philosophical theory in moral psychology be helpful in achieving this point?

Presumably I should not be so pessimistic. Yes, there are students who are unmoved by a first-year course in current moral issues, but who cares? There are others who get it. They will change their lives, and the lives of others, in the light of what they have learned. I suppose that this is true, but there is also the dark side. Arguments have, of course, been made in favor of the importance of studying ethics. Eric Schwitzgebel (2009) offers Aristotle, Kant, Mill, Martha Nussbaum, and Michelle Moody-Adams as examples of philosophers who advocate the study of morality as a way of improving behavior. However, the study of training in moral philosophy provides reasons not only to take my skepticism seriously but also to suspect the worst: that adult moral education makes people worse, not better. James Young (1986) has pressed this case against courses in applied ethics, suggesting that they reduce students' capacities to appreciate moral issues and their complexities. Schwitzgebel and Joshua Rust have collected empirical evidence that tells against the superior moral standing of people with training in moral philosophy. We are not regarded as morally superior by our peers (Schwitzgebel and Rust 2009). Books in our field are more likely than books in other fields to go missing from libraries (Schwitzgebel 2009). This is admittedly circumstantial evidence, and responses could be made, but on the face of it these findings are telling. If we take them at face value, things do not look so good for adult moral education. The people who have it are not especially well esteemed by their peers, perhaps for good reasons.

Should we be surprised by the findings that adult moral education has, at best, intangible benefits? Not really, and the WMSH provides an explanation why this might be so. Adult moral education consists in attempts, of a variety of kinds, to get people to think about moral issues and the nature of moral values. This much goes without saying. But the vast majority of these same people are already competent moral agents, without necessarily thinking explicitly about these issues or values.

They refrain from killing, stealing, etc., for the most part and with some exceptions. They even go out of their way to pursue things that they think are *interpersonally* valuable—i.e., not only of benefit to themselves. Thus, we should think of them as antecedently competent with regard to morality. They are imperfectly but not badly sensitive to many values. We must think of their adult moral education against this backdrop: it consists in an attempt to get them to think *differently* about things to which they are already sensitive, or, at the outside, to develop sensitivity to values to which they have so far been insensitive. In terms of the distinctions presented in the previous section, education serves to facilitate or supplement environmental sensitivity. In both cases the results might be positive, but they might just as well be negative. The values to which adults are newly sensitive need not cohere well with their pre-existing sensibilities. More importantly, drawing attention to values or features of values can disrupt sensitivities already in place.

As a model, consider proficient performance in a skilled activity, such as playing a sport or a musical instrument. People come to have such skills in a wide range of ways. Formal education in such activities must proceed by drawing attention to movements, techniques, and possibilities that have so far gone unnoticed and which may interfere with the movements and techniques which someone already makes and the possibilities which someone already recognizes. Indeed, this is part of the point: when we start out we make the relevant movements poorly, so attention must be drawn to these movements in order to improve them. Interference with our tendencies is crucial for their improvement, whether we are beginners or experts. Such attention, however, can have indirect effects. We all know that, when performing a familiar activity, concentrating on one thing can interfere with how we do other things. If when playing tennis I dedicate myself to looking for chances to come to the net, my ability to hit forehands may well suffer.

I suspect that much the same goes for adult moral education. By the standards of the WMSH this is important but predictable. This view countenances a plurality of ways in which we make moral judgments, perform moral reasoning, and so on, including ways that depend greatly on features of our environment. We come to have these in a variety of ways, beginning in early childhood. Adult moral education of the sort found in university ethics courses may well supplement these in desirable

ways, but it might also disrupt them. Consider the hypothetical birds discussed in chapter 1. Teaching the west-most bird to scan for food might pay dividends for it, but it might just as well divert its attention from the movements of its flockmates, thereby interfering with its participation in the wide cognitive systems by which it already detects food. Perhaps education will eventually yield better performance overall, competency on one front sitting comfortably with competence on the other, but this is not guaranteed, and we should certainly not expect it in the short term. This goes both for our hypothetical birds and for students in university ethics classes.

This unsurprising point can be illuminated if we recall the taxonomy of types of input to systems from chapter 1. There I distinguished between input that is mediated by other agents, input that is unmediated, and dual input to which the agent has direct access and which is mediated by other agents. Training our skills of attention can, in principle, do a lot of different things to the input to our cognitive systems. It can generate new input, it can change the status of input, and it can change the sort of processing which the input undergoes. Given this array of changes, it should be no wonder that moral education can bring with it undesirable effects.

That the training of attention characteristic of moral education can result in the introduction of new input to someone's systems for moral judgment and moral reasoning is not surprising; expanding students' horizons is one of the points of such education, of course. In my introductory courses in applied ethics, I regularly include classes on moral issues connected with non-human animals, such as whether we should use such creatures for food or research. Invariably some of my students have never asked themselves these questions before. Such use of non-humans has not been an item for moral appraisal for these students. Once I draw their attention to these issues, the processes they have for moral judgment and reasoning have a new topic on which to work. It's a good thing too, both for non-humans and for me—if this were impossible, familiar sorts of moral improvement would also be impossible, and I would have seriously diminished employment opportunities.

One point of adult moral education is also to get students thinking, either in new ways or for the first time, about topics with which they are already familiar. That is to say, focusing of attention can also result in new ways of processing extant input. Consider again my first-year applied ethics

students. Many of them have opinions about abortion, but fewer of them have thought much about it. Recall the discussion of moral dumbfounding from chapter 3—for many people, judgment and reasoning about particular moral issues come apart, resulting in firm moral judgments without much to say in defense of them. When I get (cajole, force) these students to think and to write about a topic such as abortion, they are doing something to a particular object of thought that they have not done before, and they are, in effect, taking a new stance toward themselves and the world. Presumably this process morally improves some students. However, as we have seen, one of the outcomes of such reflection can be to adopt the stance of the amoralist, and perhaps even to commit oneself to a life of amoralism. Although such stances need not be morally problematic, I think that it is a stretch to see this as a desirable outcome of moral education.

The subtlest effect of drawing attention to moral issues through adult moral education is the change of status of input it can bring about. Suppose that some moral judgment and moral reasoning is accomplished via wide cognitive systems. This means that, first and foremost, for a given agent who participates in such a system, information about the topic is processed in a mediated or dual manner. Attention transforms this in two ways. The same input may be taken up in an effectively unmediated way as a person examines it without its being simultaneously shared with other agents. Moreover, both the input and the earlier processing of it can be objects of the agent's attention, again in an unmediated manner. Classroom examples include efforts by professors to get their students to reflect on the widely held status of certain ideas or on the modes of transmission of moral opinions.

It is important to note that the occurrence of one of these transformations of input does not preclude the others. The focusing of attention characteristic of adult moral education can, in principle, result in a cascade of changes to input. A student learns about a moral issue for the first time: new input. She sees links between this and other issues that she already cares about: new processing. On the basis of these realizations, she thinks critically about all the issues and the ways that she and others have been dealing with them: new inputs and new status of old inputs. On the assumptions that education and attention are not perfect and that unre-

flective, pre-adulthood beliefs, thought patterns, and practices are not necessarily problematic, we should expect this slew of changes to bring about both good and bad effects. Indeed, to the extent that such moral education consists in replacing the extended moral cognitive systems on which we normally rely with new, unpracticed individualistic ones, we should count ourselves fortunate that adult moral education ever works in a desirable manner! I say this partly in jest: extant opinions often call for reform, and disturbing the old cognitive systems seems a particularly deep way of effecting such desirable change. In some cases, two heads are better than one, so that training the one by disturbing its connection to the other will be detrimental. In other cases, too many cooks spoil the broth, such that focusing on the one yields more palatable moral soup. There is no magic bullet of moral education, wide or narrow.

Perhaps you are surprised at the spirit of the line of thought explored so far. The WMSH differs from extant work in moral psychology because of the constitutive role it gives to features of agents' environments. In contrast, psychology in general, including moral psychology, tends to be individualistic. Yet the current discussion focuses on performing moral education by changing individuals. Perhaps the lesson of the WMSH should instead be that we should make better people not by changing the people themselves but by changing the world. The novel resources for education offered by this theory are worldly ones.

In spirit I have no objection to this rejoinder. If we are interested in making the world a better place, then changing institutions, laws, practices, and more diffuse features of our social contexts strikes me as a good way to proceed. But the WMSH offers us no new options for making the world a better place. Individualistic positions already recognize opportunities for social change. Moreover, I have emphasized the importance of other agents as the resources that constitute wide moral cognitive systems. If this is correct, a practical emphasis on contextual change cannot avoid changing individual agents also.

6.11 Therapeutic Implications of the WMSH: Autism and Psychopathy

I suggested in chapter 1 that mind-reading capacities and a tendency to conform in thought are important for participation in the sorts of wide

cognitive systems that are important to moral psychology and hence can be used as clues for choosing topics when devising wide psychological hypotheses. An obvious implication is that impairments of one or both of these capacities should bring with them obstacles to participating in the wide systems in question and hence deficiencies or abnormalities of moral cognition. This is exactly what we find in autism and psychopathy. Psychopaths are notoriously and dangerously indifferent to other people. Their antisocial behavior invites the explanation, in the present context, that they suffer from some sort or sorts of deficiency in conforming their thought and behavior to patterns exhibited by the people around them. This, in turn, makes it difficult for them to participate in wide cognitive systems constitutive of normal agency. Autistic people are famously blind to the thoughts of others. The most prominent explanation of autism is that it results from and consists in mind-blindness—i.e., a failure of mind-reading capacities (see Frith 2003, especially chapter 5; Baron-Cohen 1995). The result is in one way similar to what is here hypothesized to happen to psychopaths: autistic people suffer problems with regard to participation in the wide cognitive systems to which normal moral agents have access. Yet autistic people differ from psychopaths in important ways. Psychopaths have intact mind-reading capacities (Nichols 2004, 59) yet are very dangerous to the people around them. Whatever their moral-psychological deficits, autistic people are hardly dangerous at all. In the following sections I will devise an explanation of these differences and sketch some prospects for therapy from the WMSH perspective. As with education, it will turn out that our enthusiasm should be tempered with caution in these domains.

6.12 Autism

It might seem that I am mistaken to include autistic people in a discussion of moral psychopathologies at all.[13] Autistic people are not generally thought to pose significant risks to others. Perhaps more importantly, autistic people make the moral/conventional distinction (Blair 1996; Nichols 2004).

First, should autistic people be thought of as falling inside the bounds of morality? Their relative harmlessness suggests that they should, but there are important considerations that tilt the balance in the other direction.

Often overlooked are the patients who were the subjects of Hans Asperger's seminal studies as reported in his 1944 paper "'Autistic Psychopathy' in Children." These were children with behavioral difficulties due to their autism. Although Asperger was confident that such children could and should be treated to improve their behavior, to most others they were "obnoxious brats" (Frith 1991, 7). Fritz V. ignored instructions from parents and never fit in with other children. He violated norms of politeness—e.g., by publicly playing with his own spit—and morality—e.g., by hitting other children (Asperger 1944, 39–43 in 1991 reprint). Harro L. acted in the same spirit but with different acts. Asperger records his propensity for lying and his "social unconcern" (the latter manifested in attempts at homosexual acts with other boys) (ibid., 51). Uta Frith (1991, 24–25) claims that Asperger's experience was typical: although many are loath to dwell on it, "repulsive acts" and "repugnant behavior" by autistic children are familiar to practitioners and to others who treat and live with them.

Difficult behavior is not the only way an abnormal relation to morality can manifest itself. Frith notes that many high-functioning autistics are "excessively concerned with doing the right thing" (1991, 25). "Excessive" means, of course, more than normal. Even where conduct is neither abnormally bad or good, unusual relations to morality can be found in the autistic population. Temple Grandin describes her way of thinking about morality in some detail in *Thinking in Pictures* (2006). She learned to divide social and moral rules into four categories: Really bad things, Courtesy rules, Illegal but not bad, and Sins of the system (241). This allows her to navigate social settings fairly well. Crucially, these categories serve in lieu of an understanding of the concepts of "right" and "wrong," which Grandin thinks are too abstract for autistic people to understand (240). I take it that this approach to morality—explicitly devising categories for negotiating the complex world of other humans—is not characteristic of normal moral agency. There are two ways to think of an approach such as Grandin's. One might see it as a compensating measure for a moral-psychological deficit. Alternatively, one could cast it as a distinct way—genuine, though different from more common forms—of being a moral agent. Either way, we find here a moral mind differently oriented toward both morality and the world than more typical moral minds.[14]

The second question is one of classification. I have so far spoken of "autistic people" and "high-functioning autistics," but I should be more

specific. Since the early 1990s, if not earlier, autism has been conceived of as a "spectrum disorder" (Frith 1991; Bowler 2007; Grandin 2006). This means that we should really speak of *forms of* autism. These forms share characteristics that make them all forms of the same thing, but the differences are important. Using terms from the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV), Dermot Bowler (2007, 12, 25) writes that the most common forms of autism are autistic disorder, Asperger's Disorder, and "PDD-NOS" or "pervasive developmental disorder not otherwise specified (including atypical autism)." People with Asperger's Disorder are the "high-functioning" autistics. The DSM-IV approach is the dominant one, but there are other classificatory schemas. All recognize a spectrum of disorders. Notably, all use behavioral measures to diagnose the various forms of autism (Bowler 2007, 15). The general diagnostic hallmarks are the presence of all three of the following: qualitative impairments in reciprocal social interaction, qualitative impairments in communication, and impairments in imaginative abilities. People with Asperger's Disorder share the social interaction deficits but have much better communicative and linguistic abilities. The diagnostic emphasis on behavior means that specifying just what psychological phenomena the autistic spectrum consists in, and what the differentiating causes of the various forms are, is a matter ripe for psychological hypothesizing. However, no matter what diagnostic feature impresses one most at the outset, the eventual picture will be a pluralistic one, using diverse psychological capacities to explain both the nature of autism in general and the differences between the various forms.

The implication here for wide approaches to the mind is positive yet tempered. Failure to participate in wide cognitive systems will be at most part of the story of the psychology of autistic-spectrum disorders. Whether such wide systems will be fundamental to the whole spectrum or peculiar to specific forms—or both, in the case of the involvement of multiple wide systems—is an empirical issue.

With regard to moral psychology more particularly and the WMSH especially particularly, more specific points can be made. The diagnostic hallmarks of autistic-spectrum disorders, especially deficits in social interaction, should lead us to expect fairly deep and pervasive disruption of any wide mechanisms involved in moral psychology. Since the whole spectrum is characterized by problems with social interaction, disruption of wide

systems can be hypothesized to be characteristic of autism-spectrum disorders in general. We should also expect some moral-psychological problems to be specific to particular forms of autism. Insofar as some wide moral systems require sophisticated linguistic abilities, people with autistic disorder can be expected to face problems that are not shared by Asperger's Disorder patients.

More subtly, the stakes for people with autism-spectrum disorders appear different from wide and narrow perspectives. I shall focus on the social impairments that pervade the whole spectrum. If cognitive capacities are individualistically realized, we should think of autistic people as cut off from others and from forms of social life that generally are important to people. From a wide perspective, we should see autistic people as not only cut off from others and the associated forms of social life, but also cut off from parts of their own minds. Insofar as certain sorts of interactions with others realize wide cognitive systems, autistic people are faced with cognitive obstacles as well as social ones.

Autism has proved frustrating to treat. Grandin's book *Thinking in Pictures* reads simultaneously as a vote of confidence in the many avenues of possible treatment for autistic people and as a record of hopes raised and subsequently significantly reined in. This is in keeping with the likely plurality of mechanisms involved: treating one process need not have implications for the problems due to other processes. This record of attempts provides an important perspective from which to survey wide options.

Generally, the moral-psychological problems of people with one or another autism-spectrum disorder can be treated directly or indirectly. Direct methods include trying to repair or provide a capacity which is not present, supporting capacities that are present but are impaired in some way, and developing compensating measures to deal with either absent or impaired abilities. These approaches need not be mutually exclusive. Grandin's report of her way of thinking about values is illustrative here. It is legitimate to see it as both a compensating measure and as a way of developing the sensitivity to values that Grandin finds herself to have. It is a compensating measure insofar as an explicit, linguistically encoded taxonomy of values is used for moral judgment and moral reasoning in lieu of the less explicit and (*prima facie*) less rigid sensitivities that psychologically normal people have. In a study comparing moral understanding in

autistic children with normal children and children with learning disabilities, Grant et al. (2005, 326–328) found that although autistic children made normal moral judgments, their moral justifications were typically poor and correlated with their verbal ability. Improving linguistic capacities might improve such moral reasoning skills; this seems to be what we find in Grandin's case.

Indirect methods of treatment are both subtler and, to my mind, more potent in their promise of deep change. Indirect treatment works not by targeting moral psychology directly but by addressing some other problem that happens to be related to moral-psychological capacities in important ways, such that improvement in the other area yields moral-psychological improvement. Social interaction is an obvious focal point. If there are ways of facilitating or supporting capacities for social interactivity, then the subsequent sensitivity to the minds of others and the associated forms of life should, by hypothesis, provide a platform for the wide cognitive systems the WMSH hypothesizes to be important in normal moral psychology. Notably, the measures taken to develop social interaction can be, but need not be, wide. For instance, practicing certain forms of interaction is a wide measure, insofar as it involves learning to navigate certain sorts of situations by using the resources found in interpersonal contexts. But narrow interventions are, to my mind, more promising.

Here are two examples. First, consider mind-reading abilities, which are part of the most prevalent account of the deficit responsible for autism. (See, e.g., Baron-Cohen 1995; Frith 2003.) On this account, autistic people largely lack the cognitive capacities for understanding the minds of others, whereas psychologically normal people have such capacities. This suggests a broad range of options for improving the social interaction of autistic people: improve their mind-reading capacities, either by facilitating mind reading, by supporting the limited capacities that they do have, or by developing compensating skills in lieu of normal mind reading. If such capacities are narrowly realized, then a narrow intervention could provide access to wide cognitive systems characteristic of normal moral psychology.[15] Second, autistic people often have sensory problems. It is not clear how fundamental to the autistic spectrum such problems are. Nor is it clear whether such sensory problems are due to relatively up-stream or down-stream issues.[16] In any case, sensory problems present an obvious barrier to engagement with external cognitive resources, including those realized

by other people. Addressing sensory problems can facilitate better use of the environment in general, and thereby better participation in wide cognitive systems, including moral-psychological ones.[17]

In my discussion of education, I spoke of domain-specific and domain-neutral measures. The present discussion adds some nuances to this dichotomy. The narrow interventions surveyed here are domain specific, yet they promise to provide scaffolding for wide cognitive processes that use other sorts of information and perform other psychological jobs. Such iterative power promises to extend cognitive improvements across domains, despite having a domain-specific starting point. If there is a uniquely wide contribution to be made to thought about the nature of autism and its treatment, the increased attention to such nested narrow and wide processes is a good candidate.

Although I am inclined to think that a pluralistic approach to the mechanisms at work in both moral psychology and autism is warranted, philosophical commentary on both can easily find itself pulled onto time-worn paths. Jeannette Kennett (2002) argues, in a familiarly Kantian vein, that autistic psychology shows the centrality of reason to moral psychology. Victoria McGeer (2008) contends that, as part of our moral psychology, reason is best seen as serving an affective concern for others, social stability, and cosmic order. These forms of concern are, for McGeer, the lynchpins of human moral psychology. I hope to have shown, at the very least, that a wide perspective on autism and moral psychology suggests a messier and more novel picture. Rather than broadly Kantian and Humean minds, the WMSH viewpoint on both moral psychology in general and autism in particular directs us to mechanisms aptly characterized in terms neither of reason nor of passion.

To my mind, a subtle yet notable feature of this broadening of the slate of options is the blurring of jobs traditionally associated with either reason or passion. As an example, consider a minor dispute between Victoria McGeer (2008) and Frédérique de Vignemont and Uta Frith (2008) about autistic moral psychology. McGeer suggests that the deficits that give rise to autistic-spectrum disorders yield differences in the ways that the three sorts of concern function, hypothesizing the concern for cosmic order to dominate in autistic moral psychology (2008, 253). The concerns for others' well-being and for social standing are present but different from normal. De Vignemont and Frith (2008, 278–280) contend instead that

what we find in autistic moral psychology is impairment of "allocentric" representation of the people and relationships. "Egocentric" representation represents people in direct relationship to the agent doing the representing. "Allocentric" representation represents people independently of the agent who is thinking about the people in question. In normal people, egocentric and allocentric representation are processed in deep connection with each other. De Vignemont and Frith (279) claim that in autistic people egocentric processing is more dominant, yielding "abstract" allocentric representation of themselves and others.

At first pass it might seem that we should see this as a reason-versus-passion dispute: McGeer suggests an affective problem, whereas de Vignemont and Frith suggest an information-processing problem more aptly captured by the traditional label of "reasoning." However, if we stand back from the debate we find a way to resist this interpretation. The issue in both accounts is, at least partly, a classificatory one that shows up in autistic moral judgment, reasoning, and action production. Both approaches articulate the ways in which autistic people think about and respond to others. By the standards of the WMSH, the classificatory problem probably is due in part to the way autistic people are shut out of distinctive environmental cognitive resources by their barriers to mind reading. Though there certainly are purely cognitive means of performing such classifications, I argued in chapter 2 that there could be wide affectively based systems that perform such classification. Without very clear and specific evidence for a purely Humean or Kantian mechanism, we should resist, here and elsewhere, falling into the rut of assuming that these are our most important theoretical options.

6.13 Psychopathy

Relative to autism, psychopathy is both more widely acknowledged to be a moral-psychological problem and thought to be even less promising to treat. There should be no question that psychopathy is deeply different from autism. However, in one crucial respect, the picture of psychopathy emerging from ongoing research portrays it as structurally more similar to autism than has been generally appreciated. This developing picture deeply complicates the lessons that philosophers have been inclined to draw from considerations of psychopaths.

It is not uncommon for discussions of moral psychology to mention, at greater or lesser length, "the psychopath." The psychopath is famously indifferent to moral considerations. Unlike an autistic person, the psychopath has intact mind-reading capacities (Nichols 2004). Despite this, psychopaths have uncommonly reduced emotional reactions to the suffering and well-being of others. Although some (e.g., Maibom 2005) seek an explanation of the psychopath in terms of reason, the more common view is that psychopaths, at base, suffer from an emotional deficit (Nichols 2004; Blair et al. 2005; Hauser 2006). That is, explanations of psychopathy line up along familiar Humean and Kantian lines.

The picture of psychopathy emerging from recent research complicates this approach. The 2006 *Handbook of Psychopathy*, edited by Christopher J. Patrick, is a good place to find this emerging picture. The researchers who contributed to it have lots of different interests and approach the topic from different backgrounds. One of the most important points of agreement is a general acknowledgment that the concept "psychopath" is a complex one. It stands in important relations to other phenomena, themselves subject to conceptual concerns. For instance, psychopathy is generally taken to be a subcategory of the diagnostic construct "antisocial personality disorder," about which David Lykken writes that "there is no theoretical or empirical basis for supposing that this scheme carves Nature at her joints" (2006, 4). More importantly, the construct "psychopath" promises to be usefully decomposed into more specific categories. After reviewing models of psychopathy, Ronald Blackburn concludes that "it seems unlikely that a single model will capture all [characteristics of psychopathy described in the literature] or encompass the more significant empirical findings that different models claim in support" (2006, 53). Thomas Widiger (2006, 167) and Christopher Patrick (2006, 615–617) think that the most productive work in the future will entail "dismantling" (2006, 167) the concept into more specific subcategories. One of the more common subtyping approaches to psychopathy differentiates between primary and secondary varieties. Crudely put, the primary psychopath has a constitutive affective deficit, whereas the secondary psychopath has an affective deficit due to anomalies in early learning. Further refinement of the primary/secondary taxonomy is likely to come along etiological, trait-based, and/or behavioral lines, according to Norman Poythress and Jennifer Skeem (2006, 175). Other subtaxonomies might be generated by

examining comorbidities and generating specific varieties of psychopathy in terms of other conditions. For example, C. Murphy and J. Vess (2003) suggest four subcategories: sadistic psychopaths, narcissistic psychopaths, antisocial psychopaths, and borderline psychopaths. (For a discussion, see Poythress and Skeem 2006, 177–179.) If this general sort of dismantling research is borne out, as I expect it will be, then the concept of "psychopathy" will be structurally similar to that of "autism": a label for a spectrum of conditions that share important features while differing in other significant ways.

The lesson for my present purposes is the same as it was with autism: We find ourselves in territory that is notably ripe for hypothesis formation. We should expect plural mechanisms at work in psychopathy. Some will probably be common to all forms of psychopathy, whereas others will be specific to particular subvarieties. We can expect wide mechanisms to be at most part of the picture. Whether they will be fundamental to all forms of psychopathy or not is an empirical issue.[18]

In chapter 1, I offered mind-reading capacities and mechanisms that deliver conformity with others as particularly important for participation in the wide cognitive systems hypothesized to be a central part of normal moral psychology. Autistic people suffer from barriers to mind reading; psychopaths do not. This leaves conforming mechanisms as a beckoning focal point for WMSH investigations into the problems characteristic of psychopathy.

There are three broad categories of problems that may beset conforming mechanisms. First, there are motivational problems. Those who fit into the psychopathy class may suffer impairments of motivation either to conform regarding specific topics or to conform to the views of others in general. People with problems of this sort are at best unmoved to conform to the views of their conspecifics; they may even be aversive to such conformity. Second, psychopaths may have difficulty tracking the views of others. Again, this may be domain specific or it may concern the views of others in general. Finally, psychopaths may be impaired with regard to the use of the views of others in their own psychology. As with the other broad categories, problems of this sort come in domain-specific and domain-neutral varieties. There is no reason to think that these are exclusive problems. Indeed, depending on developmental processes, they may tend to cluster.

Is there any evidence that psychopaths suffer from any of these suggested impairments? Let's begin with the motivational problems. Ever since the pioneering work of H. M. Cleckley (1941), psychopathy has been centrally associated with impoverished emotions. Of particular importance are emotions having to do with the suffering of others. Psychopaths are, virtually definitionally, *unmoved* by other people.[19] This goes not only for the behavior of others but also for their expressed emotions and judgments about their condition. Hence, it is empirically plausible to think that psychopaths suffer from motivational problems that may impair participation in the wide cognitive systems here hypothesized to be part of normal moral agency. Are these motivational impairments domain-specific, or are they domain-neutral? Insofar as I have emphasized emotional reactions to the suffering of others, there is reason to take the scope of the motivational obstacle to conforming to others to be relatively limited.

It is particularly tempting to appeal to motivational problems to explain the difference between psychopathy and autism-spectrum disorders. For instance, Frith (2003, 111–112) distinguishes between intentional empathy and sympathy (which is also called "instinctive" empathy): the first requires the ability to understand others' mental states, but the second does not. Sympathy or instinctive empathy is essentially an automatic emotional response to the suffering of others. Autistic people have such reactions, but they are deficient with regard to emotional responses that require mind reading. Frith (113–114) hypothesizes that psychopaths are deficient in sympathy despite their mind-reading capacities.[20] She surmises that it is the deficit in instinctive sympathy that renders psychopaths insensitive to the moral/conventional distinction. One way of interpreting this problem is that, on the basis of an emotional deficit, psychopaths are moved to conform their judgments and behavior neither to the views of others concerning their condition, as evidenced by their plight and emotional expressions, nor to others' views about how people ought to behave. Whereas autistic people are still sensitive to others, psychopaths suffer from a motivational impairment that all but shuts them out from the cognitive resources offered by others.

However, the emotional deficits of psychopaths need not be interpreted as posing solely a motivational obstacle that interferes with conforming to others. In chapter 2, I suggested that emotions may be open to calibration that enables them to perform complex sorts of categorization. I

hypothesized that such categorization could be accomplished if emotional initiation pathways were capable of being realized by mechanisms that track such external cognitive resources as other people. This possibility offers substantially different interpretations of the effects of the emotional poverty of psychopaths. It also brings us to the other two possible impairments to mechanisms for conforming views to those of others. Instead of constituting a motivational barrier to conforming to the views of others, the alternatives are that early-in-life emotional problems yield deficits in the tracking or the processing of the views of others, or in both. Instead of seeing psychopaths as being unmoved by the suffering of others, perhaps we should see them as, literally, insensitive to it. The view of emotions utilized here follows Prinz's perceptual theory. Accordingly, on the tracking interpretation of the emotional deficits of psychopathy, we should see psychopaths as blind to others, and hence to the cognitive resources offered by others. Alternatively, on the processing interpretation, perhaps the problem is not one of upstream sensitivity but rather one of more downstream comprehension. The possibility offered here is that psychopaths can notice the suffering of others but are incapable of "getting it." Either way, regardless of motivation, psychopaths, owing to fundamental emotional problems, would be incapable of participating in wide cognitive systems that use the views of other people.

These sorts of interpretations of the conforming problems of psychopaths derive some support from research on attention.[21] Joseph Newman and colleagues have devised a "poor response modulation" account of psychopathy based on their studies of attention (Hiatt et al. 2004; Vitale et al.2005, Hiatt and Newman 2006; Vitale et al. 2007, Glass and Newman 2009). Essentially, this view holds that, because of attention problems, psychopaths (specifically, Caucasian, low-anxiety psychopaths) are excessively focused on themselves and their own goals, to the detriment of others. Psychopaths turn out to have difficulty processing information that is secondary to their primary focus. Normal people use information about other people and their contexts to modify their pursuit of their own goals—this is "response modulation." But psychopaths have difficulty incorporating this secondary information into their judgments, their reasoning, and their activities. In technical terms, they have poor response modulation.[22]

It should be clear how deficits of this sort would compromise participation in wide cognitive systems. Suppose that there are such systems.

Though it might be the case that we participate in them by directing our attention intentionally at them, this strikes me as the exception rather than the rule. Instead, participating in such systems will often require us to process information that falls outside of our primary attentional focus. This is exactly what psychopaths are poor at doing. To the extent that such systems are important for normal moral agency, psychopaths will be effectively shut out.

Hiatt and Newman (2006, 345–346) note the contextual sensitivity of the attentional problems of psychopaths. This might make it seem as if they have domain-specific drawbacks, but in fact their problems seem to be domain-neutral. If topic X falls within the primary attentional focus of a psychopath, the psychopath will process it much as psychologically normal people do. But if the same topic falls outside of the primary concerns of this psychopath, the psychopath will have trouble processing information about that topic, regardless of what it is.

The scope of our tendencies to conform, and hence of our mechanisms that achieve such conformity, should be amenable to empirical investigation. Further development of the Asch-style experimental protocol ought to shed light on this issue. It could be the case that we have domain-specific mechanisms for conformity; the presence or absence of one may imply nothing about the presence or absence of others. Since, so far as I know, such research has not been performed, I merely note the issue and the empirical possibilities.

Does any of this tell us anything about the prospects for treating psychopathy? Yes, but the news is not good. Psychopathy is notoriously impervious to treatment. In his 1993 book *Without Conscience*, Robert Hare lamented that it seemed that nothing could be done to treat psychopaths. Nothing has changed since. Grant Harris and Marnie Rice conclude their 2006 review of treatment options by claiming that none are effective. Interestingly, Harris and Rice think that behavioral interventions offer the most hope. This is striking in view of the present focus on how moral psychology depends on worldly resources, as behavior takes place in the wide world in a way not shared by most forms of thought. The lack of effective treatment options is echoed by Michael Seto and Vernon Quinsey (2006, 590).

Instead of offering new hope, the present discussion offers a diagnosis. Suppose that, according to the present hypothesis, psychopaths face one or more obstacles to conforming their view of the world to that of others,

and that because of this they are obstructed from participating in the wide cognitive systems that are typical of normal moral psychology. Now consider the general form of treatment. One way or another, treatment measures aim at getting psychopaths to think and act more like the rest of us. If I am right that psychopathy derives, at least partly, from obstacles to conforming to others, then treatment of psychopathy faces a particularly uphill challenge. Barring deep neural change, treatment may even be doomed to fail. The reason is that reforming measures presuppose some degree of openness to conforming one's thought and action to conform to the views of others. If psychopaths are deficient with regard to conforming, then most treatment measures are bound to be inefficacious.

The treatment of psychopathy looks even more grim when we add some complexities. First, consider psychopathy from an externalist perspective. Generally, we can change people by changing their intrinsic properties or by changing their context. But a deficit in conforming to the views of others undermines both strategies, as an openness to conformity is, by hypothesis, an important condition for wide psychological processes. Thus, the point made in the preceding paragraph applies equally to efforts that attempt to change psychopaths directly and to measures that attempt to change the contexts of psychopaths in the hope of changing psychopaths indirectly. Second, the likely plurality of mechanisms responsible for different varieties of psychopathy makes matters still worse. Suppose that someone is a psychopath as a result of a variety of impairments. A particular intervention might help on one front but not on others. Even worse, as we saw in the discussion of education, a therapeutic measure might disturb compensating processes that are already in place. This holds whether the issues are narrowly or widely realized.

Conclusion

The Wide Moral Systems Hypothesis has little practical promise for education and therapy at the moment. There are two possible reasons for this. Either the WMSH is missing something, perhaps something very important, or moral psychology is so complex that practical measures at improving those who deviate from normal mature functioning is so fraught with difficulties as to be virtually impossible to accomplish satisfactorily. I think that both reasons are apt. Since it comprises a cluster of wide hypotheses

composed of data gathered in a context that is biased in favor of individualistic interpretations of human psychology, the WMSH is without doubt incomplete. At the same time, the complexity of the data suffices to establish the heterogeneity of this aspect of our minds. In view of this complexity, a complete theory of moral psychology might not offer much practical promise either.

Since so much turns on empirical work yet to be done, I cannot in clear conscience venture a confident verdict about the relative width of the moral mind. I suspect that if you find some of my more specific hypotheses about wide cognitive mechanisms tempting, you will be also tempted to see our moral psychology as relatively widely realized. Insofar as you prefer narrow explanations of particular phenomena, you should probably see normal human moral psychology as relatively narrowly realized. If at the very least the development and empirical assessment of wide hypotheses now seems worthwhile, then my job is done: the tendency to make individualistic assumptions about moral psychology has been loosened, opening the door to the proper empirical assessment of externalist hypotheses.

Notes

Chapter 1

1. In principle, the deep/shallow distinction applies not only to hypotheses but also to explanations, models, theories, and any other explanatory devices that characterize the principled discourse of science.

2. For externalist considerations against the duplication of environmental cognitive resources coupled with consideration of the evolution of human minds, see Rowlands 1999—especially the discussion of the super-beaver in chapter 4.

3. Neil Levy (2007, 58–59) also emphasizes the radical possibilities of externalism.

4. I am sympathetic to Levy's suggestion that the real interest of externalism lies in its radical implications, and that these are best seen precisely when we give up on the comparison of wide and narrow cognitive processes as invited by the Parity Principle (2007, 58–59). This would cut off Sprevak's argument at the very beginning. I agree that our pre-theoretical judgments about what counts as cognitive stand as no principled constraint on psychological theorizing. I adhere to no such constraint in the remainder of this book. For present purposes it is important to see that, even if we constrain externalism by the Parity Principle, Sprevak's argument does not work.

5. On this brief way of characterizing a system, see pp. 3–4 of Baron-Cohen 2003.

6. For animated examples, discussion, and references, see http://www.red3d.com/cwr/boids/.

7. Some on-line boid simulations allow observers to add obstacles, food, and predators, and to tinker with speed and environmental attentiveness.

8. For discussion, see Allen, Bekoff, and Lauder 1998.

9. This may be too specific a way of identifying the nature of the relevant function. The function may be the tracking of environmental threats and opportunities in general. So: if you follow your neighbors, you are in better position to find food

that they have seen but that you have not, and you are more likely to avoid predators that they have seen but that you have not.

10. To the extent that capacities can become functionally integrated in other ways, other kinds of evidence and reasoning will be relevant to determining whether wide systems are in play.

11. Bird social life, of course, turns out to be surprisingly complex and nuanced. Details can be found in Bridget Stutchbury's books (2010a,b).

12. I mean this in an explanation-neutral sense of "mind reading." There has been considerable recent philosophical and psychological debate about just what this consists in. Do we have an innate theory of mind that we use to interpret the psychology of others? Do we instead run our own minds off line, simulating the perspectives of others from the inside in order to understand them? Both? I will try to avoid committing to a particular explanation of how we understand the thoughts of others.

13. This line of thought is an important part of Simon Baron-Cohen's account of mind-reading problems in autism, to which I will return in chapter 6. For this account, and for references to important work on social aspects of human evolution, see Baron-Cohen 1995.

Chapter 2

1. There is a very important recent body of research on our abilities to draw the distinction between moral and conventional issues. I present this tradition in chapter 3 because it concerns moral reasoning more than moral judgment.

2. In a recent study (2007), Hauser and colleagues directly assessed whether moral judgments exhibit features that are not found in conscious moral reasoning. Specifically, they found that moral judgments about hypothetical cases were sensitive to the principle of double effect, but that subjects' explicit justifications of their judgments were insufficient to account for such sensitivity.

3. The interpersonal view of moral reasoning presented with Haidt gets more attention in chapter 3.

4. Or both, presumably.

5. Indeed, in chapter 4 I offer such capacities as a unifying thread through the various ways by which we attribute moral responsibility. This approach is structurally analogous to the one proposed here.

6. For critical discussion of Dancy's distinction, see McKeever and Ridge 2006.

7. I will discuss the work of Smetana and Turiel in chapter 3.

8. This is not to deny that emotions might have locationally wide realizations. This is an empirical possibility, to be assessed in the usual ways. Moreover, it is worth noting the phenomenon of emotional contagion, in which one person's displays of emotion cause others to experience the same feelings. This is not what Haidt, or perhaps anyone, means by speaking of the externalization of emotions.

9. Pluralist accounts of moral judgment are emerging. See, e.g., Greene et al. 2004 and Cushman et al. 2006.

10. I take William Casebeer's worries about prioritizing the production of representations by judgments to be in the spirit of these remarks. Casebeer (2008, 21–22) suspects that processes of coping should be more primary to our theories of moral understanding than the production of representations, and that the assumption that freestanding representations undergird coping processes is dubious.

11. Nichols and Mallon (2006, 531) cite P. Foot (1967), W. Quinn (1989), and J. J. Thomson (1976) as representative philosophers.

12. The methodological assumption here is that first-person reports can be straightforward sources of evidence about psychological mechanisms. This is a dubious assumption, but I will not question it here. The point is that, by evidential standards accepted by the research program on which Nichols and Mallon rely—standards presupposed by their experimental design—rules are not taken to be a necessary aspect of the psychology of moral judgment. To assume that appeals to welfare must implicitly rely on a rule about the wrongness of causing harm is to beg the question against ethical particularism, which rejects the idea that rules are a necessary part of morality. For discussions, see Dancy 2004 and McKeever and Ridge 2006.

13. The person-situation debate concerns the mechanisms that produce action and their dependence on or independence from contextual features. This debate is discussed in detail in chapter 5.

14. Mind reading comes in simple and complex varieties. Though there may be multiple mechanisms for achieving social conformity, differing in the sort of mind reading that they involve, *prima facie* social conformity requires mind reading no more complex than that required for drawing the moral/conventional distinction. Recognizing conventional rules requires being able to track the beliefs of others about certain kinds of interaction. Early facility with the moral/conventional distinction seems to appear during the third year of life (Nichols 2004a, 9–10).

15. From this perspective, a question that arguably is more important than how we make such judgments is why some judgments spread through given populations. The study of the epidemiology of beliefs is most relevant here; see, e.g., Sperber 1996, Boyer 2000, and Nichols 2004a. For exploration of some limits of Sperber's research program with regard to culture, see Sneddon 2003.

16. Bioethicists have noted and discussed this. See, e.g., Rachels 1975; Battin 1991.

17. The mind-reading task faces a challenge of *epistemic* width: others' ideas about values can show up in utterances, facial expressions, and actions of diverse kinds. But one kind of thing is being tracked via these clues. In contrast, the objection to the rule-following explanation is that the posited mechanism must track a *constitutively* wide class: the class of things that count as rules is so diverse as to raise problems for the idea that a single sort of mechanism tracks them.

18. For what it's worth, I take the idea that emotions can track norms to be a specific version of the position on the nature of emotions recently presented by Jesse Prinz (2004). Prinz argues that emotions are perceptions of bodily changes and, via these, of "core relational themes" (2004, 224–225). "Core" relational themes are, roughly, relations an individual has to his or her environment that pertain to that individual's welfare (ibid., 15–16). Fear, shame, and embarrassment are all discussed by Prinz as examples of emotions that fit his schema. For example, he describe shame as "a sense of unwelcome attention that occurs when one has committed a transgression that will disappoint others" (ibid., 156). This is exactly the sort of psychological job that the conformity hypothesis requires to account for the teacup and trolley cases.

Chapter 3

1. *The Moral Judgment of the Child* is quite readable; readers seeking more detail about Piaget are recommended to see his original discussion. For a summary more detailed than what is presented here, yet which shares the present attention to social interaction, see chapter 7 of Flanagan 1991.

2. Kohlberg's work is discussed in many places. Flanagan (1991) offers a useful account.

3. Susan Dwyer (2003) suggests that moral/conventional studies can be used to fill in the details of a Strawsonian account of moral responsibility. Saxe's (2005) sampling and discussion of present-day work in empirical moral psychology uses the work of Turiel (1983) and R. James Blair (1995, 1996, 1997) on the moral/conventional distinction as one of three central traditions of study of moral thought, which attests to the increasing prominence of these studies.

4. The volume of representative articles collected by Joshua Knobe and Shaun Nichols (2008) provides a good introduction to the field. Knobe and Nichols's introduction explains this method of doing philosophy and points of contention about its strengths and weaknesses. See also the special issue of *Philosophical Psychology* devoted to experimental philosophy (volume 23, 2010, number 3).

5. Although the idea that philosophical appeals to intuition are often cast in terms of what everyone believes, or of what is pre-theoretically evident or accepted, is

sometimes disputed, see Nahmias et al. 2005, 563–564 for examples of philosophers' taking this approach.

6. This line of thought applies just as much to appeal to moral/conventional distinction studies as it does to the work of experimental philosophers.

7. For important treatments, see Harman 1977 and Sturgeon 1984. For a useful discussion, see A. Miller 2003 .

8. Neil Levy (2006) makes a similar point in a defense of Haidt's position against criticisms made by Fine (2006).

9. See also Mills and Keil 2004; Keil, Rozenblit, and Mills 2004.

10. The concrete/abstract distinction has been used in recent studies of attributions of moral responsibility—e.g., Nichols and Knobe 2007.

11. The person-situation debate was sparked by Walter Mischel's review of the literature on personality and action production in *Personality and Assessment* (1968). John Doris (2002) presents and develops the implications of this debate for philosophical thinking about the virtues. A view of the debate from the perspective of personality psychology can be found in Funder 1999. Ross and Nisbett 1991 is a well-known view of the debate from the situationist side. Famous experiments in this debate are Milgram's studies of obedience (1963), the Stanford prison study (Haney et al. 1973), and Darley and Batson's "From Jerusalem to Jericho" (1973). I discuss this material in chapter 5.

12. Haidt et al. 1993; Shweder et al.1997; Rozin et al. 1999.

13. For a useful characterization of these tests, see Blair et al. 2005, 57–59. See also Smetana 2006, 122–125.

14. For precise results, see the appendix to Knobe and Roedder 2009.

15. Qualifications similar to those discussed in chapter 2 for the first social-sensitivity hypothesis apply here, but I will not discuss them.

Chapter 4

1. The work of John Martin Fischer (1994) and Mark Ravizza (Fischer and Ravizza 1998) is perhaps most closely tied to Strawson among present-day theories of responsibility. See also Wallace 1994.

2. For a recent, concise presentation of Strawson's arguments that is finer-grained than the present discussion, see McKenna 2005. Incidentally, McKenna thinks that there is a trend in recent discussions of moral responsibility that rests on a misreading of Strawson's position. The trend is a tendency to explain being morally responsible in terms of the conditions under which it is legitimate to hold people morally

responsible (ibid., 170–172). My 2005 paper is included in this trend, for better or worse. However, the present discussion is explicitly about the psychology by which we attribute responsibility, so I hope that it sidesteps questions about the worth of the trend against which McKenna argues.

3. More recently Prinz has argued that emotions are "quasi-modular" rather than modular *tout court* (2006b). There are more similarities than differences between modularity and quasi-modularity. Given both this and the greater interest in full-blown generality, I will focus on Prinz's earlier position rather than his later one.

4. Incidentally, finding this would show that the kind of information processing in question is not informationally encapsulated. Informational encapsulation is a particularly important feature of classic, vertical modules (Karmiloff-Smith 1992, 2).

5. The working assumption here is that veridical expressions of fear are desirable for this test, and that the startle eyeblink is an objective measure of such veridicality. If such veridical expressions not a necessary feature of this means of testing emotional modularity, then measurements of startle eyeblink magnitude are not necessary.

6. The follow studies are cited in Davidson and Irwin 1999: Morris et al. 1996; Morris et al. 1998; Breiter et al. 1996; Phillips et al. 1997; Whalen et al. 1998.

7. See chapter 5 for a presentation of Hurley's use of studies of visual perception by Ivo Kohler.

8. I focus on the amygdala here because it is examined by both studies, whereas only Ochsner et al. discuss the MOFC. The same thing would have to be shown for the MOFC also.

9. Wilson does not offer an explicit argument for this schema. However, given that it is a descriptive taxonomy, its very use provides an implicit argument: the schema is vindicated to the extent that it actually carves nature at the joints, and it lacks support to the extent that it fails to do this. Accordingly, I shall offer no argument for this schema other than using it to describe ways we attribute moral responsibility. To the extent that this schema omits or obscures important details, I have failed. But to the extent that the taxonomy proves to be useful, Wilson's schema finds support in the present discussion.

10. If future studies of empathy reveal a horizontally modular structure rather than a vertical one, this finding will complicate the picture of reflex reactive attitudes: they too will turn out not to be constitutively distinct from their modes of expression. Since I have already discussed empathy, I shall leave this issue aside for now.

11. The reader may have noticed a *prima facie* tension between this section's discussion of decoupling of feelings and their expressions and the discussion of horizontal modularity, which gives a constitutive role to expressions in feelings. Certainly,

findings of *widespread full* horizontal modularity for emotions would provide reason for us to think again about the possibility of decoupled expressions and feelings. However, widespread abbreviated horizontal modularity would provide no such reason. Moreover, decoupling probably would still be empirically supported, but in terms of interaction between horizontal modules. Recall Hurley's discussion of rationality and imitation.

12. See chapter 3 for experimental philosophers' findings on other topics.

13. Here I have subtracted some subtleties regarding how the experimenter's assessed their subjects' belief in the possibility of such a computer, as they are not necessary for the present discussion.

14. As with the selection from Nahmias and colleagues, I have subtracted some subtleties used to assess subjects' acceptance of determinism or indeterminism about our world.

15. Michael McKenna makes much of this aspect of Strawson's argument. He thinks that the deep insight of Strawson's position is its drawing of our attention to the importance of the "quality of will" that characterizes the motivation behind the action for which responsibility is being attributed (2005, 172–173).

16. Other interesting candidates for targets (as well as triggers and warrants) for the reactive attitudes that *might* admit of important non-mentalistic characterizations are habits and patterns. I take it that the operation of the calibration files is central to explaining how such things can be perceived, since they are not dated particulars that a person encounters.

17. Or at patterns in their actions.

18. Gopnik et al. (1999, chapter 2) and Bloom (2000, chapter 3) review empirical studies of children's mind-reading abilities. Nichols and Stich (2003, 95–96) give some attention to Meltzoff's study.

19. In chapter 5, I will develop the idea of locationally wide action-production systems.

20. I will argue for this in chapter 5.

Chapter 5

1. I will address the wider history of such studies (as presented in the appendix to C. Miller 2003) later in the chapter.

2. It is worth noting an argument to the contrary. Tom Hurka (2006) argues that everyday moral thought applies virtue terms primarily to actions and derivatively to people.

3. Doris (2002, chapter 2) offers a different schema for understanding the situationist challenge.

4. On the very idea of "moral behavior," see section 5.7 below.

5. For other reasons to be cautious about using Blasi's study as a solid foundation for resting action on thought, especially reasoning, see Haidt 2001, 823.

6. In section 5.9, I will examine ways in which this could happen. Under both (what I call) CAPS and Wide CAPS, the class of psychological units responsible for behavior includes items that are not introspectively accessible. Under Wide CAPS, the theoretical possibility is opened that there are extra obstacles to first-person access to environmental aspects of the systems that produce behavior, compared to aspects of these systems realized within the physical boundaries of agents. Given such first-person removal from the springs of action, we should treat our experiences of agency, and the predictions made on the basis of such experiences, as fallible. Empirical studies should be able to reveal this fallibility with surprising results. And they do.

7. For developmental concerns about testing for abilities in abstraction from contexts in which those abilities are actually used, see Gedeon Deák's (2006) discussion of tests that assess children's abilities to distinguish between appearance and reality.

8. The Hartshorne-May studies are not dismissed, so they still pose a problem for Sabini and Silver.

9. This way of speaking is not peculiar to situationist psychology. In fact, it is quite common. For example, Prinz begins his account of moral judgment with a description of moral emotions, and these, he thinks, arise "in the context of morally relevant conduct" (2007, 68). Such usage means that the present point has implications beyond the present discussion of situationism and the production of action.

10. For more on naturalistic and artificial contexts, see Doris 2002, 97–100.

11. References are given in the standard *Akademie* pagination. I am using James Ellington's translation.

12. James Montmarquet (2003, 359), although a friend of virtue, is skeptical of this response.

13. Sabini and Silver's more familiar description of action as produced by beliefs, desires, and values is a more specific version of this absolutely general claim.

14. Davidson (1963, 689) acknowledges a variety of "emotions, sentiments, moods, motives, passions, and hungers" that can play this role.

15. On the debate about externalism and self-knowledge, see Wilson 2003.

16. Rob Wilson (2006, 126–128) has recently made the similar suggestion that, for perception, externalist positions are more attractive the more one takes into account the "in-the-world functional role" of organisms and perceptual systems.

17. For similar reflections, see Wilson 2006.

18. For an extended discussion of Gibsonian accounts of perception, see chapter 5 of Rowlands 1999. For externalist reflections on perception, see Wilson 2006.

19. For more discussion, see Harman 1999.

20. For an extended discussion of cases, see chapter 9 of Hurley 1998.

Chapter 6

1. Unfortunately, the same term—'externalism'—is used for different ideas in both metaethics and philosophical psychology. Happily for me, the positions defended in this book are best thought of as externalist in both senses. Readers are advised to try to keep in mind the differences between these sorts of externalism and the nuances that characterize my positions.

2. For an argument in support of externalism, see Sneddon 2009.

3. David Brink (1989) has urged philosophers to see the internalism/externalism debate about the possibility of amoralism as linked to the question of whether morality is rationally required of us, but other interlocutors in this discussion have not heeded his call.

4. For a discussion of relations between domains of inquiry, see Sneddon 2004.

5. Plausible moral issues could be multiplied indefinitely here; none of them would make a difference to the nature of the chemical processes involved.

6. It might well go for other sorts of intellectual pursuit as well. Since delimiting the apparently amoral kinds of academic activity from "moralist" ones strikes me as a subtle task requiring lots of work, I think it would take us too far from the present subject. I also think that the worth of such an endeavor is far from obvious.

7. In a brief discussion of conscience and the neuroscience of morality, Patricia Churchland (2007) points to imagination, self-control, and mind-world coupling.

8. Strictly speaking, perhaps the clearest case of productive amoralism is non-animate activity. The rolling of a rock down a hill and the moving of a plant in response to sunlight are events that are not properly assessable in moral terms. They are amoral. However, I doubt that there is much to be learned about any sort of amoralism from such cases, so I shall leave them aside.

9. So far as I know, in practice no one is tempted to think that the philosopher who adopts the moral skeptic's stance, which is *necessarily receptively* amoral, is thereby excepted from moral assessment.

10. This even goes for stances as discussed by Daniel Dennett (e.g., 1987). If it weren't the case, we wouldn't be capable of contrasting, e.g., the same beings seen from the design and intentional stances respectively.

11. Ronald Milo (1983) explicitly distinguishes a form of amoralism that implies immorality from one that does not.

12. Is there such a thing as first-order attention? How can we even conceive of such a capacity? How would we distinguish first-order, domain-specific focusing of the mind internal to vision from higher-order domain neutral attention applied to vision? The answer, it seems to me, might be found in the neurological details: even if they don't seem different from the inside, maybe there are distinct neural systems for such capacities. Still, would we have discovered that we have both first-order and higher-order capacities for focusing the mind, or would we have discovered that higher-order attention is intra-personally multiply realized? I don't know, but I don't think that the answers matter for present purposes. They might matter, however, for development and implementation of possibilities for education and therapy that go beyond the bounds of the present discussion.

13. McGeer (2008) also treats autism as worthy of special attention in the study of moral psychology.

14. On Grandin's categories, see McGeer 2008, especially pp. 242–245. McGeer favors seeing autistics as having distinct yet genuine moral agency (ibid., 246–247).

15. Mind-reading capacities, of course, might be themselves widely realized. I will not explore this line of thought here. For more details, see chapter 4.

16. For a discussion, see chapter 6 of Bowler 2007. Grandin emphasizes such problems as obstacles for autistic people and laments that they are somewhat overlooked in existing treatment practices.

17. For an argument that locates sensory issues at the core of social interaction problems faced by autistic people, see Peterson and Siegal 2000.

18. For an exploration of psychopathy and antisocial behavior that diagnoses a painful individual case by appeal to multiple mechanisms, see Oakley 2007, especially the final chapter.

19. For the definition of psychopathy by emotional dysfunction, see the first chapter of Blair et al. 2005.

20. Although Frith does not say so explicitly, psychopaths also seem to lack intentional empathy, despite their mind-reading capacities. Implicit here is the idea that in normal human development instinctive sympathy is necessary but not sufficient for intentional empathy.

21. I shall be silent about whether this research calls for an interpretation in terms of a failure of upstream sensitivity or downstream processing.

22. For a review of this work, see Hiatt and Newman 2006.

References

Adams, F., and Aizawa, K. 2001. The bounds of cognition. *Philosophical Psychology* 14 (1): 43–64.

Adams, F., and Aizawa, K. 2008. *The Bounds of Cognition*. Blackwell.

Allen, C., Bekoff, M., and Lauder, G. 1998. *Nature's Purposes: Analyses of Function and Design in Biology*. MIT Press.

American Psychiatric Association. 1994. *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition. American Psychiatric Association.

Anscombe, G. E. M. 1958. Modern moral philosophy. *Philosophy* 33: 1–19. Reprinted in *Virtue Ethics*, ed. R. Crisp and M. Slote. Oxford University Press, 1997.

Aristotle. 1962. *Nicomachean Ethics*. Macmillan.

Asch, S. E. 1951. Effects of group pressures upon the modification and distortion of judgment. In *Groups, Leadership, and Men*, ed. H. Guetzkow. Carnegie Press.

Asch, S. E. 1952. *Social Psychology*. Prentice-Hall.

Asch, S. E. 1955. Opinions and social pressure. *Scientific American* 193 (5): 31–35.

Asch, S. E. 1956. Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs* 70 (9), whole no. 416.

Asperger, H. 1944 [1991]. 'Autistic psychopathy' in childhood. In *Autism and Asperger Syndrome*, ed. U. Frith. Cambridge University Press.

Athanassoulis, N. 2000. A response to Harman: Virtue ethics and character traits. *Proceedings of the Aristotelian Society* 100 (2): 215–221.

Baron, J. 1995. Myside bias in thinking about abortion. *Thinking & Reasoning* 1: 221–235.

Baron-Cohen, S. 1995. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press.

Baron-Cohen, S. 2003. *The Essential Difference: The Truth About the Male and Female Brain*. Basic Books.

Batson, C. D., Coke, S. J., Chard, F., Smith, D., and Taliaferro, A. 1979. Generality of the 'glow of goodwill': Effects of mood on helping and information acquisition. *Social Psychology Quarterly* 42: 176–179.

Battin, M. P. 1991. Euthanasia: The way we do it, the way they do it. *Journal of Pain and Symptom Management* 6 (5): 298–305. Revised 2001.

Berthoz, S., Armony, J. L., Blair, R. J. R., and Dolan, R. J. 2002. An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain* 125: 1696–1708.

Blackburn, R. 2006. Other theoretical models of psychopathy. In *Handbook of Psychopathy*, ed. C. Patrick. Guilford.

Blackburn, S. 1998. *Ruling Passions*. Oxford University Press.

Blair, R. J. R. 1995. A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57: 1–29.

Blair, R. J. R. 1996. Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders* 26: 571–579.

Blair, R. J. R. 1997. Affect and the moral-conventional distinction. *Journal of Moral Education* 26 (2): 187–196.

Blair, R., Mitchell, D., and Blair, K. 2005. *The Psychopath: Emotion and the Brain*. Blackwell.

Blasi, A. 1980. Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin* 88: 1–45.

Blevins, G., and Murphy, T. 1974. Feeling good and helping: Further phone booth findings. *Psychological Reports* 34: 326.

Bloom, P. 2000. *How Children Learn the Meanings of Words*. MIT Press.

Bowler, D. 2007. *Autism Spectrum Disorders: Psychological Theory and Research*. Wiley.

Boyd, R., and Richerson, P. J. 2005. *The Origin and Evolution of Culture*. Oxford University Press.

Boyer, P. 2000. Evolution of the modern mind and the origins of culture: Religious concepts as a limiting case. In *Evolution and the Human Mind*, ed. P. Carruthers and A. Chamberlain. Cambridge University Press.

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., et al. 1996. Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* 17: 875–887.

Brink, D. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge University Press.

Brooks, R. A. 1991. Intelligence with representation. *Artificial Intelligence* 47: 139–160.

Burge, T. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4 (1): 73–122.

Casebeer, W. 2008. Processes and moral emotions. In *Moral Psychology*, volume 3: *The Neuroscience of Morality*, ed. W. Sinnott-Armstrong. MIT Press.

Chen, S., Schechter, D., and Chaiken, S. 1996. Getting at the truth or getting along: Accuracy- versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology* 71: 262–275.

Churchland, P. S. 2007. Neuroscience: Reflections on the neural basis of morality. In *Defining Right and Wrong in Brain Science: Essential Readings in Neuroethics*, ed. W. Glannon. Dana.

Clark, A. 1997. *Being There*. MIT Press.

Clark, A. 2006. Material Symbols. *Philosophical Psychology* 19 (3): 291–307.

Clark, A. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press.

Clark, A., and Chalmers, D. 1998. The extended mind. *Analysis* 58: 10–23.

Cleckley, H. M. 1941. *The Mask of Sanity*, fourth edition. Mosby.

Clowes, R. W., and Morse, A. F. 2005. Scaffolding cognition with words. In *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, ed. L. Berthouze et al. Lund University.

Cushman, F., and Young, L. 2009. The psychology of dilemmas and the philosophy of morality. *Ethical Theory and Moral Practice* 12 (1): 9–24.

Cushman, F., Young, L., and Hauser, M. 2006. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science* 17 (12): 1082–1089.

Dancy, J. 2004. *Ethics Without Principles*. Oxford University Press.

Darley, J., and Batson, C. D. 1973. From Jerusalem to Jericho: A study of situational and dispositional variables in helping behaviour. *Journal of Personality and Social Psychology* 27: 100–108.

Darley, J. M., and Berscheid, E. 1967. Increased liking as a result of anticipation of personal contact. *Human Relations* 20: 29–40.

Darley, J. M., and Latané, B. 1968. Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* 8: 377–383.

Davidson, D. 1963. Actions, reasons, and causes. *Journal of Philosophy* 60 (23): 685–700.

Davidson, R. J., and Irwin, W. 1999. The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Sciences* 3 (1): 11–21.

Deák, G. 2006. Do children really confuse appearance and reality? *Trends in Cognitive Sciences* 10 (12): 546–550.

Dehaene, S. 1997. *The Number Sense*. Oxford University Press.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., and Tviskin, S. 1999. Sources of mathematical thinking: Behavioral and brain imaging evidence. *Science* 284: 970–974.

Dennett, D. C. 1987. *The Intentional Stance*. MIT Press.

Dent, N. J. H. 1984. *The Moral Psychology of the Virtues*. Cambridge University Press.

de Sousa, R. 1987. *The Rationality of Emotions*. MIT Press.

de Vignemont, F., and Frith, U. 2008. Autism, morality, and empathy. In *Moral Psychology*, volume 3: *The Neuroscience of Morality*, ed. W. Sinnott-Armstrong. MIT Press.

de Vignemont, F., and Singer, T. 2006. The empathic brain: How, when and why? *Trends in Cognitive Sciences* 10 (10): 435–441.

Doris, J. 1998. Persons, situations, and virtue ethics. *Noûs* 32 (4): 504–530.

Doris, J. 2002. *Lack of Character*. Cambridge University Press.

Dwyer, S. J. 2003. Moral responsibility and moral development. *Monist* 86: 181–199.

Fine, C. 2006. Is the emotional dog wagging its rational tail, or chasing it? Reason in moral judgment. *Philosophical Explorations* 8 (9): 83–98.

Fischer, J. M. 1994. *The Metaphysics of Free Will: An Essay on Control*. Blackwell.

Fischer, J. M. 1999. Recent work on moral responsibility. *Ethics* 110 (October): 93–139.

Fischer, J. M., and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Flanagan, O. 1991. *Varieties of Moral Personality: Ethics and Psychological Realism*. Harvard University Press.

Fodor, J. 1983. *The Modularity of Mind*. MIT Press.

Foot, P. 1967. The problem of abortion and the doctrine of the double effect. *Oxford Review* 5: 5–15.

Fowles, D. C. 1980. The three arousal model: Implications of Grady two-factor learning theory for heart rate, electrodermal activity and psychopathy. *Psychophysiology* 17: 87–104.

Fowles, D. C. 1988. Psychophysiology and psychopathy: A motivational approach. *Psychophysiology* 25, 373–391.

Freedman, B. 1987. Equipoise and the ethics of clinical research. *New England Journal of Medicine* 317: 141–145.

Frith, U. 1991. Asperger and his syndrome. In *Autism and Asperger Syndrome*, ed. U. Frith. Cambridge University Press.

Frith, U. 2003. *Autism: Explaining the Enigma*, second edition. Blackwell.

Funder, D. C. 1999. *Personality Judgment*. Academic Press.

Gallese, V., Keysers, C., and Rizzolatti, G. 2004. A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8 (9): 396–403.

Garner, R. 1994. *Beyond Morality*. Temple University Press.

Gazzaniga, M. S. 2005. *The Ethical Brain: The Science of Our Moral Dilemmas*. Ecco.

Gigerenzer, G. 2008. Why heuristics work. *Perspectives on Psychological Science* 3: 20–29.

Gilligan, C. 1992. *In a Difference Voice: Psychological Theory and Women's Development*. Harvard University Press.

Glannon, W. 2004. *Biomedical Ethics*. Oxford University Press.

Glass, S. J., and Newman, J. P. 2009. Emotion processing in the criminal psychopath: The role of attention in emotion-facilitated memory. *Journal of Abnormal Psychology* 118 (1): 229–234.

Goldberg, L. R. 1993. The structure of phenotypic personality traits. *American Psychologist* 48: 26–34.

Goldstick, D. 2006. Beliefs, desires and moral realism. *Philosophy* 81 (01): 153–160.

Goodale, M. A. 2001. Why vision is more than seeing. In *Naturalism, Evolution and Intentionality*. *Canadian Journal of Philosophy,* Supplementary Volume 27.

Gopnik, A., Meltzoff, A., and Kuhl, P. 1999. *The Scientist in the Crib: What Early Learning Tells Us About the Mind*. HarperCollins.

Gordon, R. 1996. Sympathy, simulation, and the impartial spectator. In *Minds and Morals: Essays on Ethics and Cognitive Science*, ed. L. May, M. Friedman, and A. Clark. MIT Press.

Grandin, T. 2006. *Thinking in Pictures: My Life with Autism*, expanded edition. Vintage Books.

Grant, C. M., Boucher, J., Riggs, K. J., and Grayson, A. 2005. Moral understanding in children with autism. *Autism* 9 (3): 317–331.

Greene, J. D., and Haidt, J. 2002. How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6 (12): 517–523.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44 (2): 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. 2001. An fMRI nvestigation of emotional engagement in moral judgment. *Science* 293 (September): 2105–2108.

Haidt, J. 2001. The emotional dog and its rational tail: A social-intuitionist approach to moral judgment. *Psychological Review* 108 (4): 814–834.

Haidt, J., Koller, S., and Dias, M. 1993. Affect, culture and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65: 613–628.

Haney, C., Banks, W. C., and Zimbardo, P. G. 1973. Study of prisoners and guards in a simulated prison. *Naval Research Reviews* 9: 1–17.

Hare, R. M. 1981. *Moral Thinking*. Clarendon.

Hare, R. 1993. *Without Conscience: The Disturbing World of the Psychopaths Among Us*. Simon and Schuster.

Harman, G. 1977. *The Nature of Morality: An Introduction to Ethics*. Oxford University Press.

Harman, G. 1999. Moral philosophy meets social psychology. *Proceedings of the Aristotelian Society* 99: 315–332.

Harman, G. 2000. The nonexistence of character traits. *Proceedings-of-the-Aristotelian-Society* 100: 223–226.

Harris, G. T., and Rice, M. E. 2006. Treatment of psychopathy: A review of empirical findings. In *Handbook of Psychopathy*, ed. C. Patrick. Guilford.

Hartshorne, H., and May, M. A. 1928. *Studies in the Nature of Character I: Studies in Deceit*. Macmillan.

Hauser, M. C. 2006. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. Ecco.

Hauser, M. C., Cushman, F., Young, L., Kang-Xing Jin, R., and Mikhail, J. 2007. A dissociation between moral judgments and justifications. *Mind & Language* 22 (1): 1–21.

Hauser, M. C., Young, L., and Cushman, F. 2008. Reviving Rawls's linguistic analogy: Operative principles and the causal structure of moral actions. In *Moral Psychology*, volume 2: *The Cognitive Science of Morality*, ed. W. Sinnott-Armstrong. MIT Press.

Hiatt, K. D., and Newman, J. P. 2006. Understanding psychopathy: The cognitive side. In *Handbook of Psychopathy*, ed. C. Patrick. Guilford.

Hiatt, K. D., Schmitt, W. A., and Newman, J. P. 2004. Stroop tasks reveal abnormal selective attention in psychopathic offenders. *Neuropsychology* 18 (1): 50–59.

Hume, D. 1740 [1978]. *A Treatise of Human Nature*. Oxford University Press.

Hurka, T. 2006. Virtuous acts, virtuous dispositions. *Analysis* 66: 69–76.

Hurley, S. L. 1998. *Consciousness in Action*. Harvard University Press.

Hursthouse, R. 1997. Virtue theory and abortion. In *Virtue Ethics*, ed. R. Crisp and M. Slote. Oxford University Press.

Hutchins, E. 1995. *Cognition in the Wild*. MIT Press.

Isen, A. M., and Levin, P. F. 1972. Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology* 21: 384–388.

Kant, I. 1785 [1993]. *Grounding for the Metaphysics of Morals*. Hackett.

Karmiloff-Smith, A. 1992. *Beyond Modularity*. MIT Press.

Kauppinen, A. 2007. The rise and fall of experimental philosophy. *Philosophical Explorations* 10 (2): 95–118.

Keil, F. C., Rozenblit, L., and Mills, C. M. 2004. What lies beneath? Understanding the limits of understanding. In *Thinking and Seeing: Visual Metacognition in Adults and Children*, ed. D. Levin. MIT Press.

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., and Fessler, D. M. T. 2007. Harm, affect, and the moral/conventional distinction. *Mind & Language* 22 (2): 117–131.

Kennett, J. 2002. Autism, empathy and moral agency. *Philosophical Quarterly* 52 (208) :340–357.

Kennett, J. 2006. Do psychopaths really threaten moral rationalism? *Philosophical Explorations* 9 (1): 69–82.

Keysers, C., Kohler, E., Umiltà, M. A., Nanetti, L., Fogassi, L., and Gallese, V. 2003. Audiovisual mirror neurons and action recognition. *Experimental Brain Research* 153 (4): 628–636.

Knobe, J. 2003. Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology* 16 (2): 309–325.

Knobe, J. 2006. The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture* 6 (1–2): 113–132.

Knobe, J., and Doris, J. 2010. Strawsonian variations: Folk morality and the search for a unified theory. In *The Moral Psychology Handbook*, ed. J. Doris. Oxford University Press.

Knobe, J., and Nichols, S. 2008. *Experimental Philosophy*. Oxford University Press.

Knobe, J., and Roedder, E. 2009. The ordinary concept of valuing. *Philosophical Issues* 19 (1): 131–147.

Kohlberg, L. 1981. *The Philosophy of Moral Development: Moral Stages and the Idea of Justice*. Harper & Row.

Kohlberg, L. 1984. *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. HarperCollins.

Kring, A. M., and Bachorowski, J.-A. 1999. Emotions and psychopathology. *Cognition and Emotion* 13 (5): 575–599.

Kupperman, J. 2001. The indispensability of character. *Philosophy* 76: 239–250.

Lave, J. 1988. *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge University Press.

Levin, P., and Isen, A. 1975. Further studies on the effect of feeling good on helping. *Sociometry* 38, 141–147.

Levitt, S. D., and Dubner, S. J. 2005. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. HarperCollins.

Levy, N. 2006. The wisdom of the pack. *Philosophical Explorations* 8 (9): 99–103.

Lykken, D. T. 2006. Psychopathic personality: The scope of the problem. In *Handbook of Psychopathy*, ed. C. Patrick. Guilford.

Maibom, H. L. 2005. Moral unreason: The case of psychopathy. *Mind & Language* 20 (2): 237–257.

Mallon, R. 2008. Reviving Rawls's linguistic analogy inside and out. In *Moral Psychology*, volume 2: *The Cognitive Science of Morality*, ed. W. Sinnott-Armstrong. MIT Press.

McCrae, R. R., and Costa, P. T., Jr. 1996. Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In *The Five-Factor Model of Personality: Theoretical Perspectives*, ed. J. Wiggins. Guilford.

McDowell, J. 1997. Virtue and reason. In *Virtue Ethics*, ed. R. Crisp and M. Slote. Oxford University Press.

McGeer, V. 2008. Varieties of moral agency: Lessons from autism (and psychopathy). In *Moral Psychology*, volume 3: *The Neuroscience of Morality*, ed. W. Sinnott-Armstrong. MIT Press.

McKeever, S., and Ridge, M. 2006. *Principled Ethics: Generalism as a Regulative Ideal*. Oxford University Press.

McKenna, M. 2005. Where Frankfurt and Strawson meet. *Midwest Studies in Philosophy* 29: 163–180.

Meltzoff, A. 1995. Understanding the intentions of others: re-enactment of intended acts by 18-month old children. *Developmental Psychology* 31: 838–850.

Merritt, M. 2000. Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice* 3: 365–383.

Milgram, S. 1963. Behavioral study of obedience. *Journal of Abnormal and Social Psychology* 67: 371–378.

Mill, J. S. 1863 [2001]. *Utilitarianism*, second edition. Hackett.

Miller, A. 2003. *An Introduction to Contemporary Metaethics*. Polity.

Miller, C. 2003. Social psychology and virtue ethics. *Journal of Ethics* 7: 365–392.

Millikan, R. G. 1996. Pushmi-pullyu representations. In *Mind and Morals: Essays on Ethics and Cognitive Science*, ed. L. May, M. Friedman, and A. Clark. MIT Press.

Mills, C. M., and Keil, F. C. 2004. Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology* 87: 1–32.

Milner, A. D., and Goodale, M. A. 1995. *The Visual Brain in Action*. Oxford University Press.

Milo, R. 1983. Amorality. *Mind* 92 (October): 481–498.

Milton, J. 2003. *Paradise Lost*. Penguin.

Mischel, W. 1968. *Personality and Assessment*. Wiley.

Mischel, W. 1999. Personality coherence and dispositions in a cognitive-affective personality systems (CAPS) approach. In *The Coherence of Personality: Social Cognitive*

Bases of Consistency, Variability, and Organization, ed. D. Cervone and Y. Shoda. Guilford.

Mischel, W., and Shoda, Y. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review* 102: 246–268.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., and Grafman, J. 2005. The neural basis of human moral cognition. *Nature Reviews Neuroscience* 6: 799–809.

Montmarquet, J. 2003. Moral character and social science research. *Philosophy* 78: 355–368.

Moody-Adams, M. M. 1997. *Fieldwork in Familiar Places*. Harvard University Press.

Morris, J. S. 1996. A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* 383: 812–815.

Morris, J. S., Friston, K. J., Büchel, C., Frith, C. D., Young, A. W., Calder, A. J., et al. 1998. A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain* 121: 42–57.

Murphy, S., Haidt, J., and Bjorklund, F. 2000. Moral dumbfounding: When intuition finds no reason. *Lund Psychological Reports* 2, 1–23.

Murphy, C., and Vess, J. 2003. Subtypes of psychopathy: Proposed differences between narcissistic, borderline, sadistic, and antisocial psychopaths. *Psychiatric Quarterly* 74: 11–29.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. 2005. Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18 (5): 561–584.

Nichols, S. 2002. Is it irrational to be amoral? How psychopaths threaten moral rationalism. *Monist* 85 (2): 285–304.

Nichols, S. 2004a. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press.

Nichols, S. 2004b. The folk psychology of free will: Fits and starts. *Mind & Language* 19 (5): 473–450.

Nichols, S. 2006. Folk intuitions about free will. *Journal of Cognition and Culture* 6: 57–86.

Nichols, S., and Knobe, J. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs* 41 (4): 663–685.

Nichols, S., and Mallon, R. 2006. Moral dilemmas and moral rules. *Cognition* 100: 530–542.

Nichols, S., and Stich, S. 2003. *Mindreading: An Integrated Account of Pretense, Self-Awareness and Understanding Other Minds*. Oxford University Press.

Nisan, M. 1987. Moral norms and social conventions: A cross-cultural comparison. *Developmental Psychology* 23: 719–725.

Nucci, L., and Turiel, E. 1993. God's word, religious rules, and their relation to Christian and Jewish children's concepts of morality. *Child Development* 64: 1475–1491.

Nussbaum, M. C. 1997. *Cultivating Humanity*. Harvard University Press.

Nussbaum, M. C. 2007. On moral progress: A response to Richard Rorty. *University of Chicago Law Review* 74: 939–960.

Oakley, B. 2007. *Evil Genes: Why Rome Fell, Hitler Rose, Enron Failed, and My Sister Stole My Mother's Boyfriend*. Prometheus Books.

Ochsner, K. N., Bunge, S. A., Gross, J. J., and Gabriele, J. D. E. 2002. Rethinking feelings: An fMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience* 14 (8): 1215–1229.

O'Neill, O. 1986. A simplified version of Kant's ethics: Perplexities of famine and world hunger. In *Matters of Life and Death: New Introductory Essays in Moral Philosophy*, second edition, ed. T. Regan. Random House.

Patrick, C. 2006. Back to the future: Cleckley as a guide to the next generation of psychopathy research. In *Handbook of Psychopathy*, ed. C. Patrick. Guilford.

Perkins, D. N., Farady, M., and Bushey, B. 1991. Everyday reasoning and the roots of intelligence. In *Informal Reasoning and Education*, ed. J. Voss, D. Perkins, and J. Segal. Erlbaum.

Peterson, C. C., and Siegal, M. 2000. Insights into theory of mind from deafness and autism. *Mind & Language* 15 (1): 123–145.

Phelps, E. 2004. The human amygdala and awareness: Interactions between emotion and cognition. In *The Cognitive Neurosciences III*, ed. M. Gazzaniga. MIT Press.

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., et al. 1997. A specific neural substrate for perceiving facial expressions of disgust. *Nature* 389: 495–498.

Piaget, J. 1932 [1965]. *The Moral Judgment of the Child*. Free Press.

Poythress, N. G., and Skeem, J. L. 2006. Disaggregating psychopathy: Where and how to look for subtypes. In *Handbook of Psychopathy*, ed. C. Patrick. Guilford.

Prinz, J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. Oxford University Press.

Prinz, J. 2006a. The emotional basis of moral judgments. *Philosophical Explorations* 9 (1): 29–43.

Prinz, J. 2006b. Is emotion a form of perception? *Canadian Journal of Philosophy. Supplementary* 32: 137–160.

Prinz, J. 2007. *The Emotional Construction of Morals*. Oxford University Press.

Putnam, H. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7: 131–193.

Quinn, W. 1989. Actions, intentions, and consequences: the doctrine of double effect. *Philosophy & Public Affairs* 18: 334–351.

Rachels, J. 1975. Active and passive euthanasia. *New England Journal of Medicine* 292: 78–80.

Reynolds, C. W. 1987. Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics* 21 (4): 25–34.

Richerson, P. J., and Boyd, R. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. University of Chicago Press.

Rizzolatti, G., and Craighero, L. 2004. The mirror-neuron system. *Annual Review of Neuroscience* 27: 169–192.

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. 1996. Premotor cortex and the recognition of motor actions. *Brain Research. Cognitive Brain Research* 3 (2): 131–141.

Ross, L., and Nisbett, R. E. 1991. *The Person and the Situation: Perspectives of Social Psychology*. Temple University Press.

Rowlands, M. 1999. *The Body in Mind: Understanding Cognitive Processes*. Cambridge University Press.

Rowlands, M. 2003. *Externalism: Putting Mind and World Back Together Again*. McGill–Queen's University Press.

Rozenblit, L., and Keil, F. C. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 92: 1–42.

Rozin, P., Lowery, L., Imada, S., and Haidt, J. 1999. The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity. *Journal of Personality and Social Psychology* 76 (4): 574–586.

Rupert, R. 2004. Challenges to the hypothesis of extended cognition. *Journal of Philosophy* 101 (8): 389–428.

Rupert, R. 2009. *Cognitive Systems and the Extended Mind*. Oxford University Press.

Ryle, G. 1949. *The Concept of Mind*. University of Chicago Press.

Sabini, J., and Silver, M. 2005. Lack of character? Situationism critiqued. *Ethics* 115 (April): 535–562.

Saxe, R. 2005. Do the right thing. *Boston Review*, September/October. Available at http://bostonreview.net.

Schafe, G. E., and LeDoux, J. E. 2004. The neural basis of fear. In *The Cognitive Neurosciences III*, ed. M. Gazzaniga. MIT Press.

Schaefer, S. M., Jackson, D. C., Davidson, R. J., Kimberg, D. Y., and Thompson-Schill, S. L. 2002. Modulation of amygdalar activity by the conscious regulation of negative emotion. *Journal of Cognitive Neuroscience* 14 (6): 913–921.

Schwitzgebel, E. 2009. Do ethicists steal more books? *Philosophical Psychology* 22 (6): 711–725.

Schwitzgebel, E., and Rust, J. 2009. The moral behaviour of ethicists: Peer opinion. *Mind* 118 (472): 1043–1059.

Seto, M. C., and Quinsey, V. L. 2006. Toward the future: Translating basic research into prevention and treatment strategies. In *Handbook of Psychopathy*, ed. C. Patrick. Guilford.

Shoda, Y., and Mischel, W. 1996. Toward a unified, intra-individual dynamic conception of personality. *Journal of Research in Personality* 30: 414–428.

Shoda, Y., and Mischel, W. 2000. Reconciling contextualism with the core assumptions of personality psychology. *European Journal of Personality* 14: 407–428.

Shoda, Y., Mischel, W., and Wright, J. C. 1994. Intraindividual stability in the organization and patterning of behavior: incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology* 67 (4): 674–687.

Shweder, R. A., and Much, N. C. Mahapatra, M., and Park, L. 1997. The 'big three' of morality (autonomy, community, divinity), and the 'big three' explanations of suffering. In Morality and Health, ed. A. Brandt and P. Rozin. Routledge.

Singer, T. 2006. The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neuroscience and Biobehavioral Reviews* 30 (6): 855–863.

Slaate, H. A. 1981. *Modern Science and the Human Condition*. University Press of America.

Slote, M. 1995. Agent-based virtue ethics. *Midwest Studies in Philosophy* 20: 83–101.

Slote, M. 2001. *Morals from Motives*. Oxford University Press.

Smetana, J. G. 1981. Preschool children's conceptions of moral and social rules. *Child Development* 52: 1333–1336.

Smetana, J. G. 1993. Understanding of social rules. In *The Child as Psychologist*, ed. M. Bennett. Harvester Wheatsheaf.

Smetana, J. G. 2006. Social-cognitive domain theory. In *Handbook of Moral Development*, ed. M. Killen and J. Smetana. Erlbaum.

Smith, C. 2003. *Moral, Believing Animals: Human Personhood and Culture*. Oxford University Press.

Smith, M. 1994. *The Moral Problem*. Blackwell.

Smith, M. 2004. *Ethics and the a priori: Selected Essays on Moral Psychology and Meta-Ethics*. Cambridge University Press.

Sneddon, A. 2003. Naturalistic study of culture. *Culture and Psychology* 9 (1): 5–29.

Sneddon, A. 2004. Prichard, Strawson, and two objections to moral sensibility theories. *Journal of Philosophical Research* 29: 289–314.

Sneddon, A. 2005. Moral responsibility: The difference of Strawson and the difference it should make. *Ethical Theory and Moral Practice* 8 (3): 239–264.

Sneddon, A. 2009. Alternative motivation: A new challenge to moral judgment internalism. *Philosophical Explorations* 12 (1): 41–53.

Sperber, D. 1996. *Explaining Culture*. Blackwell.

Sprevak, M. 2009. Extended cognition and functionalism. *Journal of Philosophy* 106: 503–527.

Sreenivasan, G. 2002. Errors about errors: Virtue theory and trait attribution. *Mind* 111 (January): 47–68.

Strawson, P. F. 1962 [1974]. Freedom and resentment. In *Freedom and Resentment and Other Essays*. Methuen.

Sturgeon, N. 1984. Moral explanations. In *Morality, Reason, and Truth*, ed. D. Copp and D. Zimmerman. Rowman and Allanheld.

Stutchbury, B. 2010a. *The Bird Detective: Investigating the Secret Lives of Birds*. HarperCollins.

Stutchbury, B. 2010b. *The Private Lives of Birds: A Scientists Reveals the Intricacies of Avian Social Life*. Walker.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *Monist* 59: 204–217.

Turiel, E. 1997. The development of morality. In *Handbook of Child Psychology*, fifth edition, ed. W. Damon and N. Eisenberg. Wiley.

Turiel, E. 1983. *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.

Turiel, E. 2006a. Thought, emotions, and social interactional processes in moral development. In *Handbook of Moral Development*, ed. M. Killen and J. Smetana. Erlbaum.

Turiel, E. 2006b. The development of morality, revised. In *Handbook of Child Psychology*, fifth edition, ed. N. Eisenberg, W. Damon, and R. Lerner. Wiley.

Turiel, E., Killen, M., and Helwig, C. 1987. Morality: its structure, functions, and vagaries. In *The Emergence of Morality in Young Children*, ed. J. Kagan and S. Lamb. University of Chicago Press.

Ungerleider, L. G., and Mishkin, M. 1982. Two cortical visual systems. In *Analysis of Visual Behavior*, ed. D. Ingle, M. Goodale, and R. Mansfield. MIT Press.

van Inwagen, P. 1992. Reply to Christopher Hill. *Analysis* 52: 56–61.

Vitale, J. E., Brinkley, C. A., Hiatt, K. D., and Newman, J. P. 2007. Abnormal selective attention in psychopathic female offenders. *Neuropsychology* 21 (3): 301–312.

Vitale, J. E., Newman, J. P., Bates, J. E., Goodnight, J., Dodge, K. A., and Petit, G. S. 2005. Deficient behavioral inhibition and anomalous selective attention in a community sample of adolescents with psychopathic and low-anxiety traits. *Journal of Abnormal Child Psychology* 33: 461–470.

Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Harvard University Press.

Wegner, D. M. 2002. *The Illusion of Conscious Will*. MIT Press.

Weyant, J., and Clark, R. D. 1977. Dimes and helping: The other side of the coin. *Personality and Social Psychology Bulletin* 3: 107–110.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., and Jenike, M. A. 1998. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience* 18: 411–418.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., and Rizzolatti, G. 2003. Both of us disgusted in *my* insula: The common neural basis of seeing and feeling disgust. *Neuron* 40 (October): 655–664.

Widiger, T. 2006. Psychopathy and DSM-IV psychopathology. In *Handbook of Psychopathy*, ed. C. Patrick. Guilford.

Williams, B. 1972. *Morality: An Introduction to Ethics*. Harper & Row.

Wilson, R. 1994. Wide computationalism. *Mind* 103 (411): 351–372.

Wilson, R. 1995. *Cartesian Psychology and Physical Minds: Individualism and the Sciences of the Mind*. Cambridge University Press.

Wilson, R. 2000. The mind beyond itself. In *Metarepresentations: A Multidisciplinary Perspective*, ed. D. Sperber. Oxford University Press.

Wilson, R. 2001. Two views of realization. *Philosophical Studies* 104 (1): 1–31.

Wilson, R. 2003. Individualism. In *Blackwell Guide to Philosophy of Mind*, ed. S. Stich and T. Warfield. Blackwell.

Wilson, R. 2004. *Boundaries of the Mind*. Cambridge University Press.

Wilson, R. 2005. *Genes and the Agents of Life*. Cambridge University Press.

Wilson, R. 2006. Critical notice of Mohan Matthen's *Seeing, Knowing, and Doing: A Philosophical Theory of Sense Perception* (Oxford, 2005). *Canadian Journal of Philosophy* 36 (March): 117–132.

Wilson, R., and Clark, A. 2008. How to situate cognition: Letting nature take its course. In *The Cambridge Handbook to Situated Cognition*, ed. P. Robbins and M. Aydede. Cambridge University Press.

Wilson, R., and Keil, F. C. 1998. The shadows and shallows of explanation. *Minds and Machines* 8: 137–159.

Young, J. O. 1986. The immorality of applied ethics. *International Journal of Applied Philosophy* 3 (2): 37–43.

Zagzebski, L. T. 1996. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge University Press.

Index